VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra telekomunikační techniky

Analýza kepstra řečových vzorků

Cepstral analysis of speech samples

2013 **Jakub Vychodil**

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra telekomunikační techniky

Zadání bakalářské práce

Student: **Jakub Vychodil**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 2601R013 Telekomunikační technika

Téma: Analýza kepstra řečových vzorků
Cepstral Analysis of Speech Samples

Zásady pro vypracování:

1. Rozbor hlasového traktu a vzniku lidské řeči
2. Rozbor segmentálních parametrů využívaných ve zpracování řeči
3. Návrh algoritmu pro extrakci základního tónu řeči pomocí kepstrálních koeficientu a Mel-frekvenčních kepstrálních koeficientu.
4. Vyhodnocení použitých metod

Seznam doporučené odborné literatury:

1. PSUTKA, Josef. Mluvíme s počítačem česky: the definitive guide. Vyd. 1. Praha: Academia, 2006, 746 s. ISBN 80-200-1309-1.
2. Blanka Heringová, Petr Hora. MATLAB Díl I. Práce s programem. http://www.cdm.cas.cz/czech/hora/vyuka/mvs/tutorial.pdf

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Pavol Partila**

Datum zadání: 16.11.2012
Datum odevzdání: 07.05.2013

prof. RNDr. Vladimír Vašinek, CSc.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlášení studenta

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

Dne: 6.5. 2013

.....
podpis studenta

Poděkování

Rád bych poděkoval panu Ing. Pavolu Partilovi za odbornou pomoc a konzultaci při vytváření této bakalářské práce.

Abstrakt

Tato bakalářská práce je tematicky zaměřena na extrahování základního tónu lidského hlasu ze zvukové nahrávky. Proces analýzy a následného vyhodnocení je prováděn pomocí tří metod a to kepstrální analýzy, analýzy Mel-frekvenčních kepstrálních koeficientů a autokorelační funkce. Jednotlivé metody využívají rozdílných postupů pro získání základní frekvence. Cílem je tedy tyto odlišnosti vyhodnotit a určit vhodnost každé metody pro určitý typ hlasového vzorku.

Klíčová slova

Kepstrum, kepstrální koeficienty, Mel-frekvenční kepstrální koeficienty, autokorelační funkce, lidská řeč, základní frekvence, analýza hlasu

Abstract

This bachelor's thesis is thematically focused on extraction of fundamental frequency of the human voice from the sound record. The process of analysis and subsequent evaluation is carried out by three methods, namely cepstral analysis, Mel-frequency cepstral coefficients analysis and autocorrelation function. Each method use different procedures for extraction of the fundamental frequency. The objective is evaluation of these differences and determines the suitability of each method for a certain type of speech sample.

Key words

Cepstrum, cepstral coefficients, Mel-frequency cepstral coefficients, autocorrelation function, human speech, fundamental frequency, voice analysis

Seznam použitých zkratek

Zkratka	Anglický význam	Český význam
AKF	Autocorrelation function	Autokorelační funkce
CC	Cepstral coefficients	Kepstrální koeficienty
FFT	Fast Fourier transformation	Rychlá Fourierova transformace
MFCC	Mel-frequency cepstral coefficients	Mel-frekvenční kepstrální koeficienty
PCM	Pulse-code modulation	Pulzní kódová modulace

Seznam obrázků

Seznam tabulek

Obsah

1 Úvod

Již od pradávna je nejpřirozenějším způsobem komunikace mezi lidmi řeč. S postupným evolučním rozvojem člověka se vyvíjela i jeho mozková kapacita a spolu s ní centrum řeči. Začala postupně vznikat slovní zásoba a první primitivní jazyk. Rozvoj člověka však nezůstal jen u mluvené formy řeči a dal tak vzniku písma v podobě piktogramů vyjadřujících celá slova. Později se slova začala skládat z písmen. V dnešní době řeč ovládá již každý jedinec, kromě němých a hluchoněmých. Ovšem s darem řeči se nenarodí nikdo. V genetickém kódu člověka je pouze zakódován předpoklad se jí naučit. Dítě se řeči učí postupně tak, jak roste a vyvíjí se. V době jeho vývoje je třeba, aby žilo v prostředí kde je mluvená řeč aktivním prostředkem pro komunikaci mezi lidmi. Řeči se učí odposloucháváním a snahou se pochopit význam jednotlivých slov, které se následně snaží napodobit. Díky evoluci se rodíme s mnohými předpoklady, kterým se musíme naučit, jako vzpřímené chůzi nebo chápavému užívání rukou. Některé vlastnosti se však rozvíjí samostatně, jako pud sebezáchovy či potřeba lidí ulehčovat si práci. Proto člověk vynalezl techniku, kterou se obklopujeme, aby měl život lehčí. Výpočetní kapacita počítačů a ostatních podobných zařízení roste geometrickou řadou každý rok. Pro naši pohodlnost využíváme nejčastěji k ovládání zařízení klávesnici, myš nebo dnes nejmodernější formu, dotykové rozhraní. Potřeba člověka usnadnit a urychlit komunikaci je stále větší proto ovládání veškeré techniky kolem nás pomocí řeči není až takovou vzácností, např. hlasové ovládání telefonu. Lidská řeč je nástroj složitý, pro jeho analýzu je potřeba složitých algoritmů. Lze díky tomuto velmi efektivně ovládat zařízení, ovšem počítač nedokáže automaticky pochopit vyslovenou větu. Nedokáže z tónu rozpoznat emoce mluvčího nebo dvojsmyslný význam věty. Vyslovená věta může být stejná, jako ta uložená, ale pokud se hlas mluvčího liší od hlasu člověka, který vzorek nahrál a uložil do paměti, pak jej počítač nepozná, ač je věta zcela stejně slovo od slova vyslovena. Někteří lidé trpí vadami řeči, například koktáním. Větu vyslovenou koktajícím člověkem počítač pochopí zcela jinak, než byla myšlena. Jsou porovnávány parametry hlasu a ne přímo význam jednotlivých slov. Obdobný problém nastává, pokud má počítač překládat psaný text do mluveného projevu, tedy Text-to-Speech. Počítač syntetizuje text do podoby lidského hlasu, ovšem výsledný projev je velmi nepřirozený. Počítač nedokáže věrohodně napodobit lidský hlas, nemůže do něj vložit žádné emoce, jelikož žádné nemá. Pauzy mezi slovy jsou stejně dlouhé, tempo je strojově přesné a tón je po celou dobu řeči neměnný. Pokud vysloví stejnou větu tisíckrát, bude vždy stejná. To je u člověka nepravděpodobné.

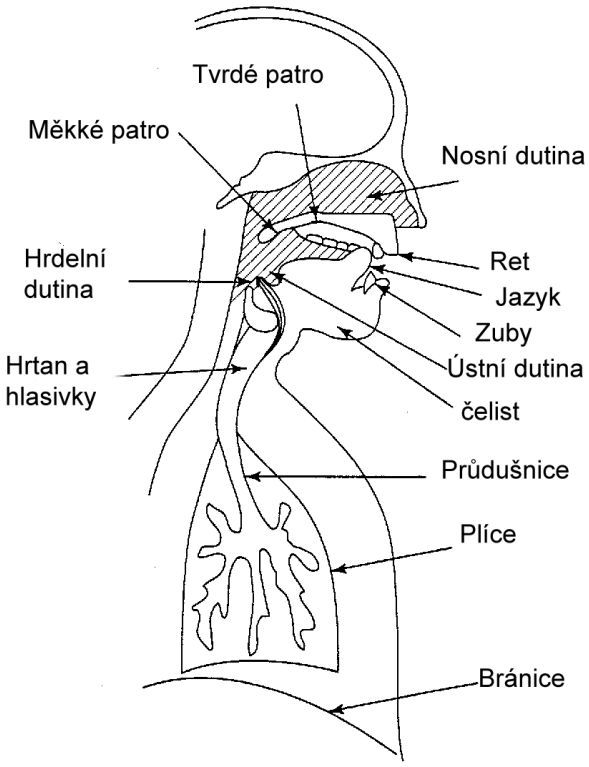
Cílem této bakalářské práce je analýza kepstra řečových vzorků pomocí metod pracujících s kepstráními koeficienty a Mel-frekvenčními kepstrálními koeficienty. Pro kvalitnější porovnání jednotlivých metod je zahrnuta i autokorelační funkce, jako třetí metoda extrakce základního tónu hlasu. Druhá kapitola se zabývá teoretickou částí objasňující fyzikální podstatu řeči a její vznik. Dále je zde popsán systém vnímání zvuku a jednotlivé části sluchovodu. Třetí kapitola obsahuje teoretický rozbor procesu předzpracování spolu s jednotlivými kroky, které je nutné před vlastní analýzou provést pro kvalitní vyhodnocení parametrů. Druhá polovina třetí kapitoly je zaměřena na samotný proces extrakce segmentových parametrů. Přičemž jsou zde teoreticky popsány jednotlivé parametry hlasu a následně rozebrány všechny tři metody analýzy. O teorii z kapitol dva a tři se opírá praktická část závěrečné práce, a to ve čtvrté kapitole. Je zde popsán algoritmus extrakce základního tónu a princip práce s grafickým rozhraním vytvořeným v návaznosti na tento algoritmus. Prostřednictvím rozhraní jsou naměřeny hodnoty pro celkové zhodnocení jednotlivých metod analýzy. Toto vyhodnocení je součástí čtvrté kapitoly. Extrahované hodnoty základní frekvence jsou rozděleny podle emočních stavů a pohlaví řečníků. Následně je statisticky vyhodnocena účinnost každého z postupů extrakce.

2 Lidská řeč

Mluvená řeč je přenášena komunikačním kanálem v podobě akustického signálu. Pod pojmem akustický signál si lze představit zvukové vlny reprezentované chvějícím se elastickým okolím ve frekvenci slyšitelného spektra. Toto okolí představuje komunikační kanál, ve kterém je řeč od úst mluvčího přenášena až k uším posluchače. Výsledná podoba mluveného projevu je závislá na rozpoložení řečníka a primárně na stavbě jeho hlasového traktu. Mimo jiné jsou parametry řeči ovlivňovány i právě charakteristikou komunikačního kanálu. Mluvené slovo nemusí být přenášeno pouze volným prostorem, ale také třeba přes metalické vedení. Řeč může být zkreslena, pokud mluvčí trpí vadou řeči nebo je jeho hlasové ústrojí poškozeno v důsledku poranění, popřípadě nemoci. [1]

2.1 Tvorba řeči

Hlasový trakt je ústrojí, ve kterém vzniká řeč. Toto ústrojí je složeno z několika orgánů, zobrazených na obr. 2.1.



Obrázek 2.1: Hlasový trakt

Tyto jednotlivé orgány ve své podstatě prvotně neplní funkci tvůrců zvuku. Jejich hlavní funkce jsou mnohem důležitější a podstatnější pro život jedince, např. vedení potravy, rozpoznávání chuti nebo dýchání. S procesem dýchání je spjata i tvorba hlasu. Při výdechu bránice tlačí na plíce. Z plic je odváděn vzduch průdušnicí přes hrtan, kde jsou umístěny hlasivky. Dále proud pokračuje hltanem do ústní dutiny, odkud vychází mezi rty ven z lidského těla. Tento proud vzduchu je nosným signálem pro výsledný hlas. Průchodem celého hlasového ústrojí je modulován do výsledné podoby. Člověk dokáže sám záměrně ovlivňovat výslednou podobu hlasu. Lze mluvit hluboce nebo naopak tón hlasu zvýšit oproti původnímu charakteru hlasového projevu, dále také můžeme sami ovlivňovat rychlost či frázování mluvy dle potřeby. To vše je pro nás oproti počítači zcela intuitivní.

V hrtanu je umístěn nejpodstatnější orgán pro tvorbu hlasu, a to hlasivky. Je to dvojice pružných svalů vedoucích napříč hrtanem v oblastní jeho nejmenšího průměru. Mohou se dle potřeby zavírat a otvírat, přičemž vytvoří průchod mezi sebou. Tento otvor se nazývá hlasivková štěrbina. Při běžném dýchání, tedy když člověk nemluví, jsou hlasivky otevřeny, aby mohl vzduch proudit dále ven z těla. Štěrbina je zcela odkrytá a utváří průchod trojúhelníkového tvaru o šířce 8mm, která netvoří vzduchu žádný odpor. Ovšem při vzniku hlasu začnou hlasivky naplňovat svou druhou funkci tzv. fonační. Svaly se stáhnou k sobě, tzn. hlasivková štěrbina je zcela uzavřena. Při výdechu proud vzduchů způsobí, že pružné hlasivky pod jeho tlakem začnou kmitat. Hlasivková štěrbina se rychle otvírá a zavírá. Vzduchový proud je segmentován na menší bloky chvějícího se vzduchu, tedy akustickou vlnu vnímanou uchem jako zvuk. Tato zvuková vlna je základním nosným akustickým signálem určujícím výšku hlasu. [1]

2.2 Základní frekvence hlasu

Při průchodu vzduchu přes hlasivky, kdy vlivem tlaku začnou kmitat, se frekvence chvění hlasivek přenáší na vzduchový tok. Tuto frekvenci pak nazýváme základní frekvenci hlasivkového tónu a je označována jako $\mathbf{F_0}$. Z této frekvence lze snadno odvodit i základní periodu hlasivkového tónu $\mathbf{T_0}$ (viz vztah 2.1).

$$T_0 = 1 / F_0 \tag{2.1}$$

Tabulka 2.1: Základní frekvence

	F_0 [Hz]
Muži	*80 – 160*
Ženy	*150 – 300*
Děti	*200 - 600*

U každého jedince se frekvence liší, jak je vidět z tab. 2.1. Muži mají hlubší hlas než ženy. Tón dětského hlasu je samozřejmě vyšší a pohlaví dítěte nehraje zásadní roli, jako u dospělých lidí. Tyto rozmezí ovšem nestanovují přesně dané rozsahy hlasů, u mužů může spodní hranice rozsahu klesnout až na hodnotu 50 Hz a to především u trénovaných pěveckých hlasů. Operní zpěvačky jsou schopny dosáhnout tónu hlasu o frekvenci až 1000Hz. [1]

2.3 Vytváření struktury tónu a šumové složky

Základní tón nesený vzduchovým tokem je při své cestě modulován. Jako poslední modulátor je považováno artikulační ústrojí složené z nadhrtanové dutiny a artikulačních orgánů. Tato skupina orgánů je schopna vytvořit širokou škálu nejrůznějších zvuků. Hlavní podíl na variabilitě má samozřejmě jazyk. Je to velice flexibilní sval, který změnou svých proporcí určuje celkový tvar ústní dutiny, a tedy i přeměnu hlasu v nejrozmanitější zvuky ve výsledku tvořící řeč. V nadhrtanových dutinách dochází také k zásadním změnám zvuku. Nadhrtanové dutiny se skládají z dutiny ústní, nosní a hrtanové. V těchto částech traktu dochází k rezonanci hlasu, zesílení energie a vzniku široké skály tónů. Tyto tóny se nazývají formanty a jsou značeny od formantu o nejmenší frekvenci $F_1,F_2,\ldots,F_n$. Formanty se dělí na hlavní a vedlejší, přičemž vedlejší jsou utvářeny v dutině hrtanové. V ústní dutině vzniká hlavní formant s frekvenčním rozsahem od 175 Hz až do 3600 Hz. Tato frekvence je úměrná právě hlavně postavení jazyka a zubů spolu se rty. [1]

2.4 Vnímání zvuku

Pro vnímání zvukových vln z okolí člověku slouží ucho. Je to orgánu složený ze tří částí a to zevního ucha, středního ucha a vnitřního ucha. Na rozdíl od řeči se člověk již rodí s darem sluchu. Je to jeden z pěti základních lidských smyslů. Díky sluchu se od narození učíme odposloucháváním řeči lidí z okolí rozeznávat různé zvuky, k nimž jsme schopni automaticky přiřadit jejich zdroj. Lidské ucho je velice citlivý orgán ukrytý z větší části uvnitř hlavy. Frekvenční rozsah zvuku vnímaného uchem také není neomezený. Člověk je schopen pomocí sluchu vnímat zvuky ve frekvenčním rozmezí od 16Hz – 20kHz. S přibývajícím věkem a také působením škodlivých vlivů z okolí, jako hluk, se tento frekvenční rozptyl zmenšuje. Proto by si člověk měl svůj sluch chránit. [1]

2.5 Zevní ucho

Jedinou viditelnou částí ucha je boltec, který je stavěn pro příjem zvuků přicházejících zepředu z okolí před posluchačem. Slouží ke směrování zvuků z okolí do zevního zvukovodu. Což je dutina s otevřeným koncem navazujícím na boltec a na svém druhém konci je uzavřená bubínkem. Zvuk je při průchodu zvukovodem zesilován v rozmezí od 3kHz do 5kHz. [1]

2.6 Střední ucho

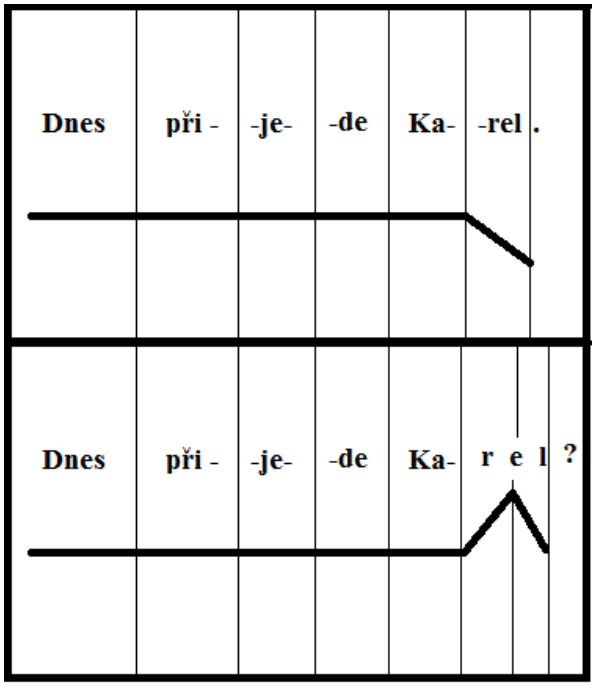
Bubínek tvoří rozhraní mezi uchem zevním a středním, které je uloženo v tzv. bubínkové dutině. Střední ucho je v soubor tří kůstek – kladívka, kovadlinky a třmínku. Zvuk dopadající na bubínek je přenášen na tyto kůstky, které jsou rozkmitávány a vibrace tím vzniklé putují dále do oválného okénka, což je přechod mezi středním a vnitřním uchem. Bubínková dutina je napojena na Eustachovu trubici. Jejím úkolem je vyrovnávat tlak ve středním uchu a přivádět vzduch z nosohltanu, se kterým je spojena svým druhým koncem. Ve středním uchu se také nachází dvojice svalů, které chrání citlivé vnitřní ucho před poškozením zvuky o nízké frekvenci. Při této situaci se svaly stáhnou a utlumí tak kmitání bubínku zhruba o 20 dB. Tyto svaly ovšem reagují s menším zpožděním. Za další stupeň ochrany lze považovat rotaci třmínku. Při vzrůstající intenzitě zvuku třmínek postupně obloukem rotuje a působí tak méně na okénko vnitřního ucha díky tomu nedochází k poškození.[1]

2.7 Tempo řeči

Je to v podstatě rychlost řečového projevu člověka. Tato rychlost je závislá na složitosti vyslovovaných slov spojených do vět. Tempo může charakterizovat různé formy projevu. Pokud je člověk například rozčilený a křičí, je síla hlasu a tempo řeči rychlejší než pokud je zcela vyrovnaný a klidný. Pokud pronášíme slavnostní projev, snažíme se k posluchačům mluvit pomalu a srozumitelně, aby každé slovo bylo jasně rozpoznatelné. Tempo řeči lze považovat za individuální znak každého jedince. Většina lidí při rozhovoru s druhým člověkem mluví normálním tempem. Rychlost je úměrná tak, aby bylo jednotlivým slovům dobře rozumět a byla zachována určitá plynulost projevu. Pokud mluvčí mluví příliš rychle, pak dochází ke snížení kvality projevu a často je nutné určitá slova nebo věty znovu opakovat, jelikož posluchač nestíhá zpracovávat kvantum informací a také často ani zcela správně neporozumí jednotlivým slovům. Většinu vlastností hlasu lze do jisté míry intuitivně a přirozeně ovlivňovat, stejně tak i tempo. Člověk většinou při rozhovoru změnu tempa mluvení nijak zvlášť nevnímá s výjimkou záměrné změny. Pro různá nářečí je rychlost řeči hlavním charakteristickým znakem, kdy je důraz kladen na jiné slabiky než při spisovné výslovnosti. Jazyky jako italština, čínština, španělština a další jsou od českého jazyka odlišné vysokým tempem a rychlými změnami intonace mluveného projevu a navenek tak působí velice chaoticky. [1][2]

2.8 Intonace

Dalším parametrem lidského hlasu je intonace. Je nositelem emocí a charakteru vyslovených slov. Změnou intonace je jedinec schopen z věty oznamovací vytvořit větu tázací. V mluvené podobě jazyka je nám charakter věty zcela jasný díky intonaci. Pro psanou formu jazyka byla pro rozlišení zavedena interpunkční znaménka – tečka, čárka, otazník a vykřičník. Pokud tedy čteme věty: „Dnes přijede Karel.“ „Dnes přijede Karel?“ První věta je oznamovací, druhá tázací a proto při čtení jejich charakter automaticky odlišíme různou intonací. Je zřejmé, že intonace se mění u poslední slabiky – ***rel*** (Obrázek 2.2). Ovšem změnu intonace, zapříčiněnou emočním rozpoložením z psané věty nijak nejde rozpoznat na první pohled. Lze tento stav jen odhadnout ze souvislostí mezi jednotlivými větami a z obsahu textu. Proto nemá intonace žádný vliv na syntaxi ani strukturu vět v psané podobě. Pro vyjádření postoje, zdůraznění jisté skutečnosti nebo jen emotivního zabarvení věty je intonace vhodný nástroj, ovšem nedílnou součástí je také tempo. Jak tempo, tak intonace jsou ovšem záležitostí mluveného proslovu. Počítač při konverzi textu do mluvené podoby změnu tempa nebo intonace sám nedokáže aplikovat. Nepotřebuje nijak vyjádřit svůj postoj nebo zdůraznit emoce, jelikož žádné nemá a nedokáže je ani vypozorovat z textu. Samozřejmě by se tento proces dal implementovat nějakým algoritmem, který by emoce rozpoznával z textu porovnáváním souvislostí a významů jednotlivých slov. Tento algoritmus by ovšem byl velice složitý a bylo třeba obrovské kvantum informací, které by porovnával, aby byl schopen vyhodnotit, jak má přizpůsobit intonaci nebo změnit rychlost mluvy. Software typu text-to-speech jsou většinou využívány nevidomými. Usnadňují jim práci s počítačem a ostatními zařízení. Strojová chladnost a monotónnost hlasu počítače je ovšem nepodstatná, důležitá je funkčnost, která nevidomému zásadně ulehčuje práci s počítačem a na jeho projevu nijak nesejde. Z textu, který zařízení překládá, může člověk sám rozpoznat emoce nebo jiné skutečnosti, které se normálně zdůrazňují změnou tónu hlasu. Je to podobné, jako bychom text sami četli.[6]

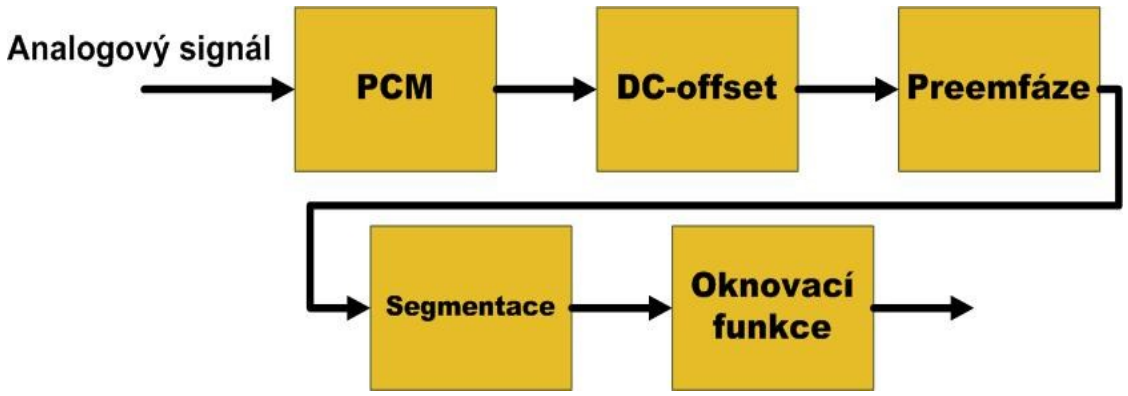


Obrázek 2.2: Srovnání intonace oznamovací a tázací věty

3 Zpracování a parametrizace hlasového signálu

3.1 Pre-processing

Před samotnou analýzou jakékoliv nahrávky hlasu je třeba tento vzorek předpřipravit. Při nahrávání řeči může docházet k zanášení nežádoucích elementů do nahrávky vlivem přechodu informace z fyzikální podoby do digitální. Těchto vlivů může být mnoho, jako stejnosměrná složka nebo útlum energie signálu průchodem prostředí. Snahou procesů zmíněných v následujících kapitolách je odstranění těchto vlivů a vhodné připravení nahrávky. Celý postup je znázorněn v podobě diagramu na obr. 3.1. Pokud by signál nebyl vhodně předpřipraven byla by celá analýza chybná z důvodu ovlivnění hodnot rušivými parametry. [10]



Obrázek 3.1: Vývojový diagram předzpracování

3.1.1 PCM

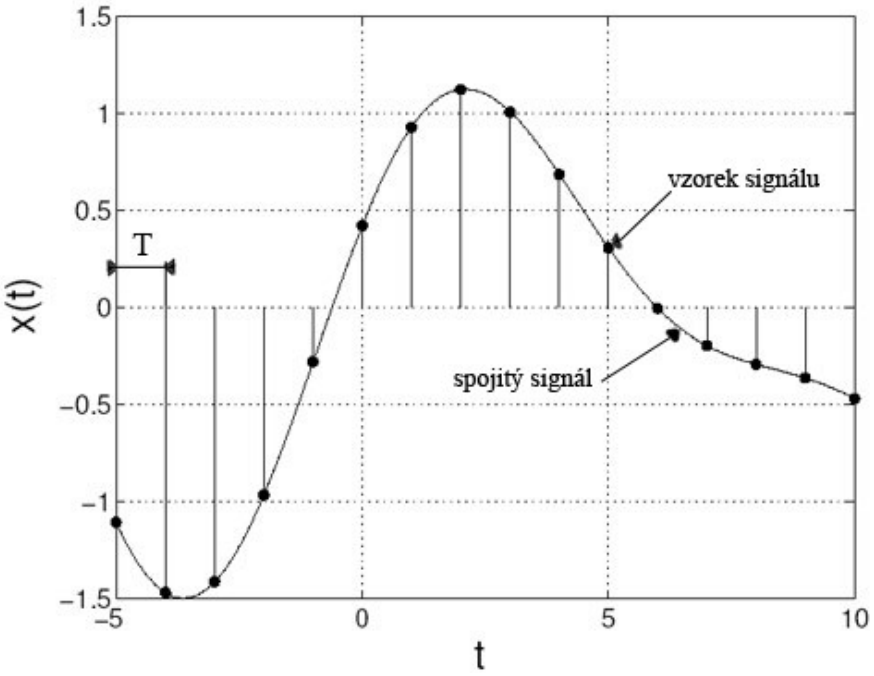
Hlas každého člověka je reprezentován jako vlnění vzduchu na určité frekvenci. Tato analogová podoba je pro lidský sluch postačující. Ovšem pro zpracování zvuku počítačem je tento formát zvuku nevhodný. Počítač pracuje s digitálními informacemi, proto je zvukový signál zaznamenán pomocí mikrofonu a je třeba jej převést na posloupnost čísel dvojkové soustavy. Tento postup převodu analogového signálu na diskrétní posloupnost se nazývá pulzní kódová modulace – **PCM**. V první fázi modulace je třeba analogový signál navzorkovat. V určitém časovém intervalu jsou odebírány vzorky ze spojitého analogového signálu. Časový interval mezi jednotlivými vzorky se nazývá perioda **T**, z níž lze jednoduše získat hodnotu vzorkovací frekvence $\mathbf{F_{vzr}}$ dle vztahu (3.1).

$$T = 1/F_{VZR} \tag{3.1}$$

Z definice Shannon-Kotělnikova teorému vyplívá, že pro zpětnou rekonstrukci signálu je třeba, aby vzorkovací frekvence $\mathbf{F_{vzr}}$ byla dvakrát větší než frekvence signálu původního. Tak je zajištěno odebírání vzorků v množství postačujícím pro rekonstrukci.

Dalším krokem modulace je kvantizace. Jednotlivým vzorkům jsou přiřazovány hodnoty, tzv. kvantizační úrovně. Rozsah těchto úrovní je zvolen tak, aby bylo možno pokrýt celý spojitý signál. Pokud hodnota vzorku, která se rovná hodnotě amplitudy spojitého signálu v čase odebrání vzorku neodpovídá hodnotě kvantizační úrovně, je jeho hodnota snížena na hodnotu nejblíže nižší

kvantizační úrovně. Tímto „ořezáváním“ vznikají kvantizační chyby, což je v podstatě rozdíl mezi původní hodnotou a hodnotou stávající daného vzorku. Počet kvantizačních chyb lze snížit zvýšením počtu kvantizačních úrovní. [3]



Obrázek 3.2: Vzorkování signálu

3.1.2 **Odstranění jednosměrné složky**

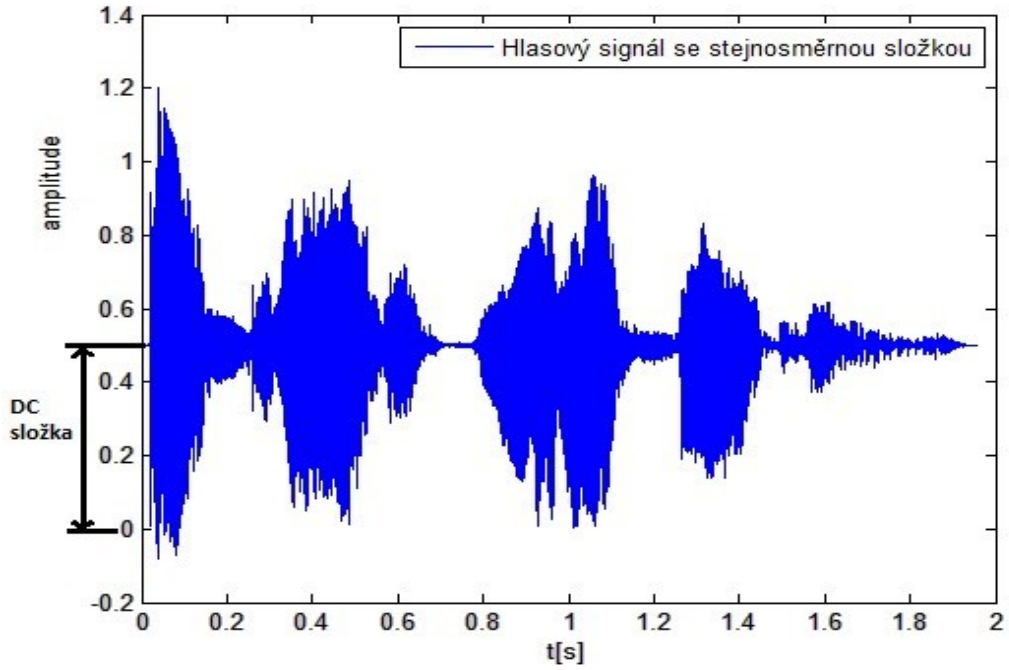
Při pořizování zvukové nahrávky často dochází k problému, že je do signálu zanesena součást stejnosměrného proudu. Tento rušivý element může ve výsledku způsobovat problémy při zpracování signálu, např. při výpočtu energie dochází k nepřesnému výpočtu její hodnoty. Po výpočtu amplituda signálu je navýšena o určitou konstantní hodnotu a signál není souměrný podle nulové osy. Je tedy nutné tuto složku odstranit.

$$\mu_s = 1/N * \sum_{n=1}^{N} s(n) \qquad (3.1)$$

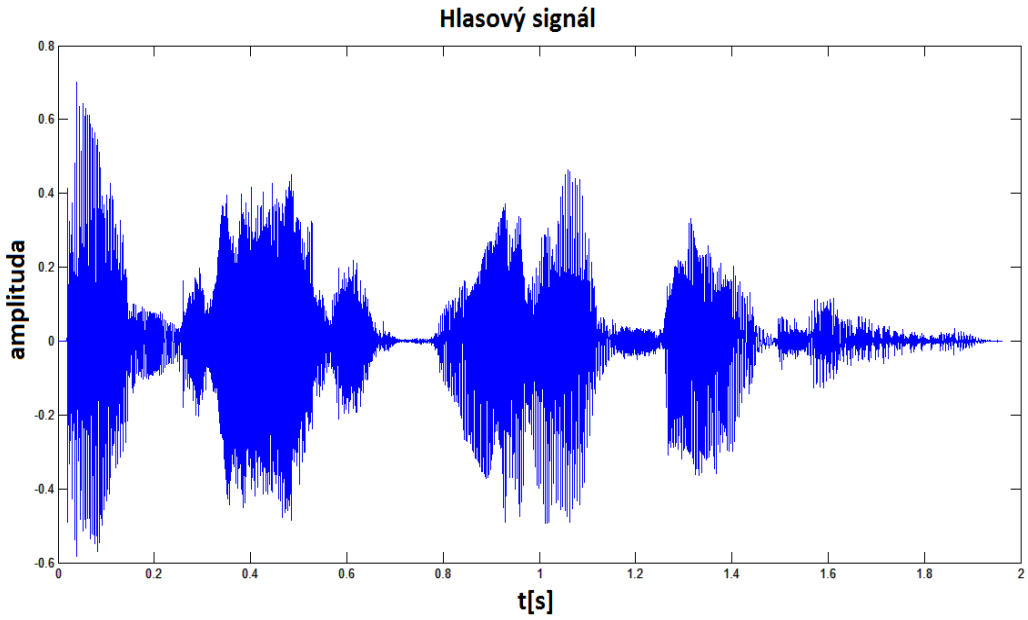
Dle vztahu 3.1 je patrné, že pro odstraněné stejnosměrné složky je třeba nejprve signál zprůměrovat. Konstanta **N** udává celkový počet vzorků řečového signálu a hodnota **n** určuje číslo aktuálního vzorku signálu.

$$s(n)' = s(n) - \mu_s \qquad (3.2)$$

Konstanta $\boldsymbol{\mu_s}$ udává velikost stejnosměrné složky a tedy i posunutí signálu od nulové osy. [10]



Obrázek 3.3: Signál se stejnosměrnou složkou

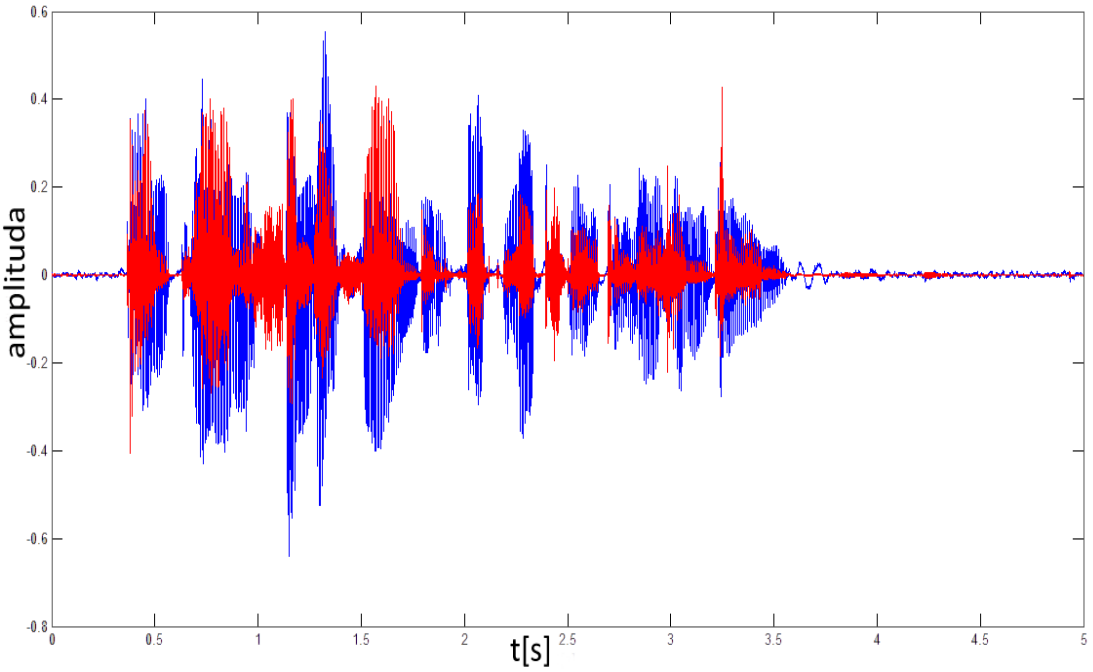


Obrázek 3.4: Vystředěný signál bez stejnosměrné složky

3.1.3 **Preemfáze**

Průchodem volným prostorem akustický signál fyzikálně reprezentující řeč je utlumován. Složky s vyšší frekvencí jsou postupně zkreslovány o 20dB/dek. Tento útlum je třeba vyrovnat filtrací. Některé složky dosahují hodnot důležitých pro analýzu právě v oblasti vyšších kmitočtů spektra. Filtrace je dána následujícím vztahem 3.3. [10] Od hodnoty vzorku signálu je odečtena hodnota vzorku předcházejícího navýšená o určitou konstantní hodnotu z rozmezí 0,9 až 1. Tímto tedy vznikne nový vyfiltrovaný vzorek a útlum je vyrovnán.

$$s(n)' = s(n) - m * s(n-1); m \in \langle 0{,}9 - 1 \rangle \tag{3.3}$$

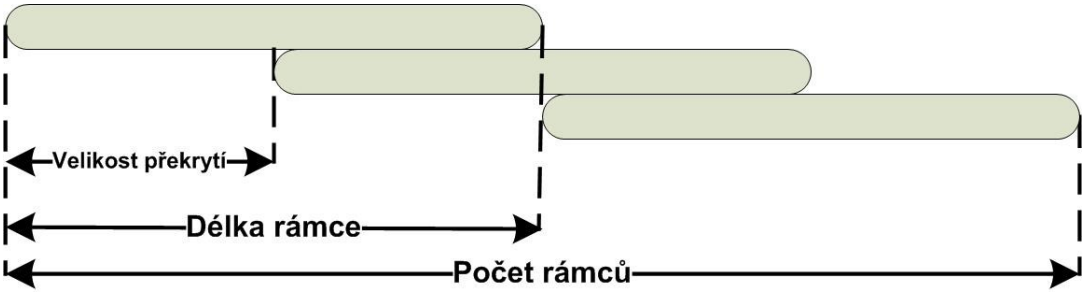


Obrázek 3.5: Průběhy signálu před preemfází a po preemfázi

Na obrázku 3.5 je znázorněn, jako modrý průběh signál před filtrací a červený průběh popisuje řečový signál po preemfázi. Je viditelné, že v oblasti vyšších frekvencí je energie navýšena. A útlum je tak vykompenzován.

3.1.4 Segmentace

Signál vytvořený hlasivkami je modulován postavením a tvarem jednotlivých hlasových orgánu. Při změně základního tónu hlasu řečové orgány změní své postavení. Ze strojového hlediska lze tuto změnu považovat za pomalou. Změna je postupná a není jednorázová, proto hlas považujeme za signál nestacionární. Nestacionarita nám definuje, že změny nejsou v čase konstantní a signál nemá vždy stejné vlastnosti. Pro práci se signálem v časové oblasti je mnohem vhodnější si jej rozdělit na menší segmenty, u nichž lze předpokládat, že jsou stacionární. Stacionarita segmentů je také závislá na vhodně zvolené délce segmentu, nejčastěji se volí délka segmentu 5 až 20ms. Pro plynulý přechod mezi segmenty je užíváno 50% překrytí segmentů.[4][5]



Obrázek 3.6: Grafické znázornění segmentace

Z hodnot délky rámce a velikosti jejich překrytí lze jednoduše získat celkový počet rámců pro segmentaci celého signálu dle vztahu 3.4.

$$N\_frame = 1 + floor\left[\left(\frac{l_{signal} - o_{sample}}{F_{vzr}}\right) / l_{frame}\right] \tag{3.4}$$

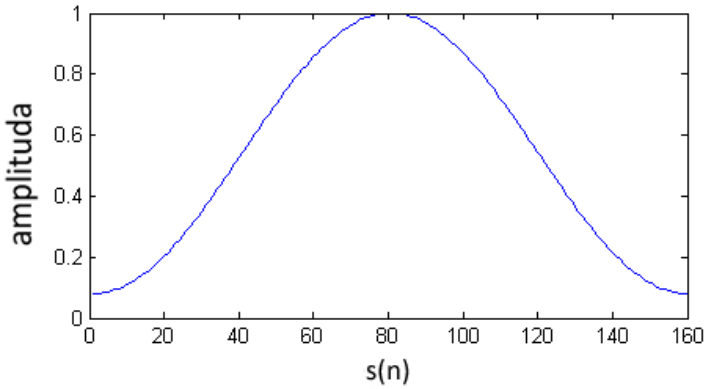
Proměnná **N_frame** ve vztahu 3.4 nese výslednou hodnotu počtu rámců pro daný signál. Funkce **floor** je funkcí používanou v Matlabu, kdy každé číslo je zaokrouhleno k nejbližší nižší celé hodnotě. V této funkci se nachází rozdíl počtu vzorků v signálu a velikost překrytí ve vzorcích dělení hodnotou vzorkovací frekvence. Proměnná **l_frame** skýtá hodnotu počtu vzorků jednoho rámce. Jelikož je počet vzorků v jednom rámci vždy zaokrouhlen na nejbližší nižší hodnotu mohlo by se tedy stát, že by některým vzorkům na konci signálu nebyl přiřazen segment. Proto je vždy k výslednému počtu segmentů připočten jeden segment navíc pro zbylé vzorky. Zbytek nevyužitého místa posledního segmentu je naplněn nulovými hodnotami.

3.1.5 Oknovací funkce

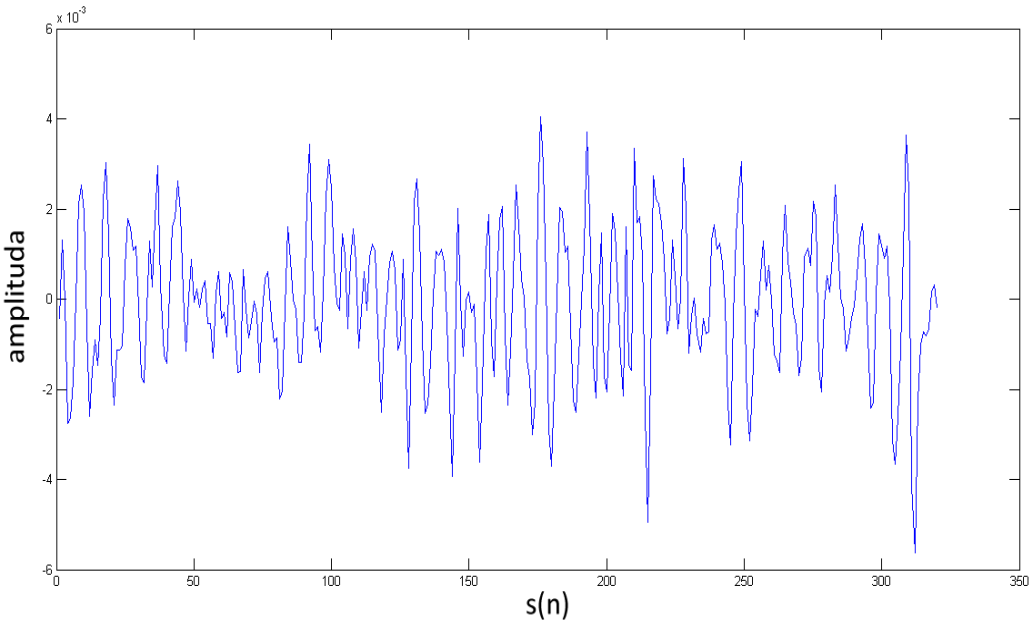
Během procesu segmentace se signál rozdělí na jednotlivé segmenty. Snahou je zajistit plynulé navázání sousedních segmentů mezi sebou prostřednictvím polovičního překrytí. Vznikají zde však ostré přechody mezi rámce během této procedury. Je třeba tuto nelinearitu kompenzovat pro zachování konstantnosti signálu z hlediska frekvence a předcházení tak vzniku chyby během zpracování. Nenulové přechody mezi segmenty jsou vyhlazeny pomocí oknovací funkce. Každý rámec vynásobíme oknovací funkcí samostatně. Funkce vybírá jednotlivě vzorky a přiřazuje jim určitou váhu. Nejčastěji se používá Hammingovo okno právě z důvodů vyhlazení okrajů.[10][11][12]

$$w(n) = 0{,}54 + 0{,}46 * \cos\left[\left(\frac{1}{2} * N - n\right)\frac{2 * \pi}{N}\right]; n \in \langle 0, N-1\rangle \tag{3.5}$$

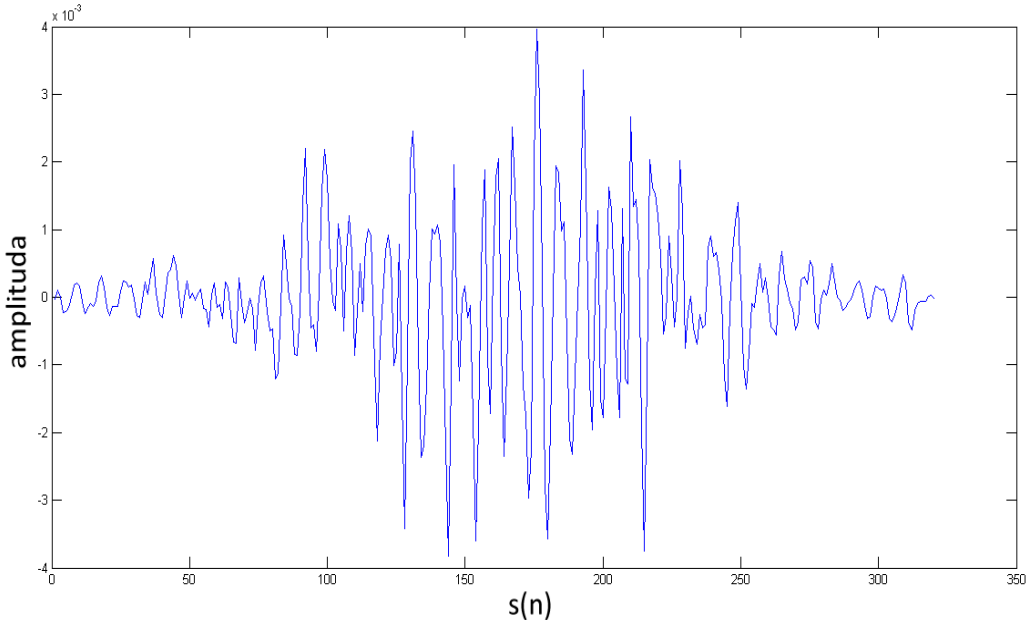
Vztah 3.5 udává funkci pro oknování signálu. Počet vzorků **N** udává šířku Hammingova okna (obr. 3.7) a proměnná **n** znamená pozici vzorku v oknu. Je třeba tedy stanovit šířku okna a následně tímto oknem vynásobit jednotlivě každý rámec signálu zvlášť.



Obrázek 3.7: Hammingovo okno



Obrázek 3.8: Signál před váhováním



Obrázek 3.9: Signál po vyvážení Hammingovým oknem

3.2 Extrakce segmentových parametrů

3.2.1 Energie lidského hlasu

Lidský hlas stejně, jako každý jiný zvuk má svoji sílu. Tuto sílu lze považovat za hlasitost, intenzitu nebo také energii signálu. Hlasitost hlasu je závislá na síle zvukové vlny produkované hlasovým traktem. Částice vzduchu se pohybují směrem od zdroje a naráží do sousedních částic, kterým předávají svou energii. Při předání energie z jedné částice na druhou, vzniká nárazem úbytek energie. Proto je zvuk slyšitelný jen na určitou vzdálenost, která závisí na jeho energii. Pokud chceme mluvit hlasitěji, hlasivky se stáhnou více k sobě a vytvoří tak malou štěrbinu. Tímto malým otvorem

projde jen malé množství vzduchu a pro jejich rozkmitání je třeba větší síla. Silným proudem vzduchu stažené hlasivky začnou vibrovat na vyšší frekvenci a zvuková vlna získá větší energii. Výsledný hlasový projev přicházející k uším posluchače se pak jeví, jako silnější, tudíž hlasitější. Hlasivky jsou namáhány více než obvykle. Člověk pak chraptí nebo nemůže mluvit vůbec. K těmto jevům dochází namáháním hlasivek. Operní pěvci svůj hlas trénují několik let. Rozšiřují si tak svůj hlasový rozsah, poté jsou schopni vyzpívat mnohem vyšší polohy hlasu nebo naopak nižší, než normální člověk a jejich hlasivky nejsou tak namáhány, ovšem mají také své meze. Hlasivky jsou prvním modulátorem zvuku v hlasovém traktu. Zbytek hlasového traktu se otevře více než při běžné mluvě, aby proud vzduchu mohl více rezonovat cestou do ústní dutiny a následně ven. Zvuk, tak dostává na své mohutnosti. Zvyšuje se jeho amplituda a energie.

Tabulka 3.1: Úrovně hluku

Zdroj	*Intenzita [W/m^2]*	*Úroveň intenzity[dB]*
Práh slyšitelnosti	*$1*10^{-12}$*	*0*
Šustění listů	*$1*10^{-11}$*	*10*
Šepot	*$1*10^{-10}$*	*20*
Běžná konverzace	*$1*10^{-6}$*	*60*
Hlučná ulice	*$1*10^{-5}$*	*70*
Vysavač	*$1*10^{-4}$*	*80*
Velký orchestr	*$6.3*10^{-3}$*	*98*
Walkman na nejvyšší hlasitost	*$1*10^{-2}$*	*100*
První řada na rockovém koncertu	*$1*10^{-1}$*	*110*
Práh bolesti	*$1*10^{1}$*	*130*
Start stíhačky	*$1*10^{22}$*	*140*
Protržení bubínku	*$1*10^{4}$*	*160*

Tabulka 3.1 ukazuje naměřené úrovně hluku v jednotlivých situacích. Energii řečového signálu je definovaná vztahem 3.6.

$$E = \frac{1}{N}\sum_{n=0}^{N-1}(s[n])^2 \tag{3.6}$$

[4][5][9]

3.2.2 **ZCR**

Zero-crossing rate je veličina určující kolikrát amplituda řečového signálu změní svou polaritu překročením nulové osy během určitého časového intervalu. Jako interval lze považovat za délku segmentu. Díky tomuto parametru a hodnotě krátkodobé energie jsou snadno rozlišitelné úseky znělé a neznělé podstatné pro analýzu vzorku a výpočet základního tónu řeči. Je třeba tyto úseky od sebe

oddělit. Neznělé úseky mají menší hodnotu ZCR a menší hodnotu krátkodobé energie, jsou neperiodické, a proto nemá význam počítat $\mathbf{F_0}$ .V těchto úsecích tedy lze předpokládat přítomnost samohlásek nebo dokonce úplného ticha. Naopak znělé části jsou periodické s vyšší krátkodobou energií, kde dochází k častým překročením nulové osy, zde jsou situovány hlásky. [9]

$$ZCR(m) = \sum |sign(s(n)) - sign(s(n-1))| \tag{3.7}$$

Změnu polarity signálu lze vypočítat pomocí rovnice 3.7. Termín **sign** značí matematickou funkci **signum**, která testuje proměnnou, zda je větší, rovna nebo menší než 0. Pokud je větší než 0 nabude funkce hodnoty 1. Je-li proměnná rovna nule, funkce se rovná nule. Pokud je menší než 0, pak bude funkce rovna -1. Ve funkci **ZCR** je tedy vzorek signálu otestován funkcí **signum** a následně je od hodnoty funkce odečtena hodnota **signum** vzorku časově předcházejícího. Pokud tedy došlo k přechodu přes nulovou osu je tento rozdíl roven dvěma.

3.2.3 Základní tón řeči

Hlavním parametrem řeči při jejím zpracování je základní tón. Je označována často jako $\mathbf{F_0}$. Jak již bylo řečeno dříve (viz kap. 2.2) je tento parametr hlasu přenášen z kmitajících hlasivek. Hodnota základní frekvence hlasu není hodnotou periodickou. Mění se v podstatě v závislosti na tom, jak řečník mluví neboli dle vrozených předpokladů jeho hlasového traktu a především tedy hlasivek. Základní tón může řečník měnit také záměrně zcela intuitivně. Pokud by změna základní frekvence hlasu byla periodická, výsledná promluva by působila nepřirozeně a poněkud strojově. K této nepřirozenosti dochází především při překladu textu na mluvenou řeč neboli text-to-speech prostřednictvím počítače. Analýzou tohoto elementu lze určit pohlaví a věk mluvčího. Dále je možné také vyčíst jeho emoční rozpoložení. Popřípadě jiné vlastnosti s hlasem spojené, jako například onemocnění dýchacích cest, jež zkresluje výrazně výslednou promluvu a jiné chyby hlasu. Metody získání základního tónu řeči lze rozdělit do několika skupin:

- Detekce v časové oblasti
- Detekce ve frekvenční oblasti
- Detekce v kepstru

[9]

3.2.4 Autokorelace

Pro detekci základního tónu lze využít autokorelační funkci, která porovná rámec signálu s jeho kopií. Je porovnáván tvar jejich signálu v časové oblasti. Po provedení autokorelace je třeba najít maximum funkce pro nalezení základního tónu. Další vrchol funkce následující za maximem nese hodnotu základní frekvence $\mathbf{F_0}$. V případě neznělého segmentu jsou hodnoty zanedbatelné, a proto pro výpočet F_0 jsou použity především znělé segmenty. Extrakce $\mathbf{F_0}$ se provede pomocí vztahu 3.8, kde **k** určuje pozici vrcholu následujícího po maximu.

$$F_0 = \frac{F_{vzr}}{k} \tag{3.8}$$

Pro oddělení znělých a neznělých rámců je třeba stanovit určitou úroveň, která tyto rámce roztřídí. Jelikož je signál nestacionární, je tedy v čase vysoce proměnlivý není možné pro celý signál nastavit jednotnou prahovou úroveň. Nejvhodnější je využít metodu centrálního klipování, která pomocí maxima rámce předcházejícího a následujícího stanovuje prahovou úroveň pro aktuální rámec.

$$P_i = \alpha \cdot \min(\max_{i-1}, \max_{i+1}) \tag{3.9}$$

Vztah 3.9 je funkcí centrálního klipování. $\mathbf{P_i}$ je prahová úroveň pro daný segment **i**. Konstanta **α** je nastavena na hodnotu 0,8. Funkce vzorky signálu ohodnocuje a normalizuje na dvojici hodnot 1,0 nebo -1. [1][9][11]

3.2.5 Kepstrální koeficienty

Řečový signál **s(n)** vzniká konvolucí budící funkce **e(n)**, přičemž je tato funkce modulována hlasovým traktem a vzniká funkce **h(n)**. Pro parametrizování celého procesu je nejvhodnější užít analýzy pomocí homomorfního systému díky němuž lze tyto složky od sebe oddělit. Mezi jeden z těchto způsobů zpracování signálů řadíme i kepstrální analýzu. Kepstrum lze definovat, jako reálnou část inverzní Fourierovy transformace logaritmu spektra signálu původního. Na vstupu celého systému je řečový signál **s(n)** vzniklý konvolucí funkce šumového signálu a funkce **h(n)** představující modulaci hlasovým traktem.

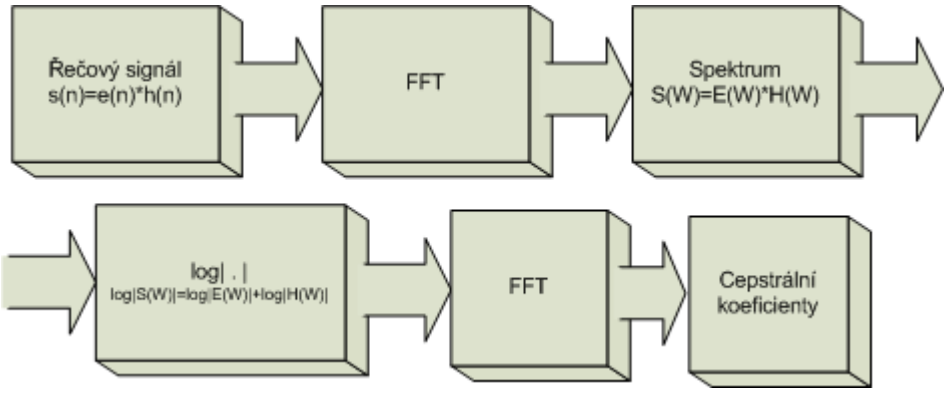
$$s(n) = e(n) * h(n) \tag{3.10}$$

Pomocí Fourierovy transformace získáme spektrum signálu **s(n)** a jeho zlogaritmováním z konvoluce vznikne součet.

$$\log|S(W)| = \log|E(W)| + \log|H(W)| \tag{3.11}$$

Provedením inverzní Fourierovy transformace získáme komplexní kepstrum, jehož reálná část je nazývána jen kepstrum.

$$c(n) = F^{-1}\{\log|S(W)|\} \tag{3.12}$$



Obrázek 3.10: Schéma systému Kepstrální analýzy

Po získání kepstra signálu lze získat základní tón hlasu, jelikož kepstrální koeficienty se v kepstru opakují po určité periodě. Tato perioda je stejná, jako perioda základního kmitočtu hlasu. [1][7][8][9]

3.2.6 Melovské kepstrální koeficienty

Každý zvuk je lidským uchem vnímán s určitým omezením. Nejvíce je zvuk vnímán v oblasti nižších frekvencí. Pak tedy lze říci, že vnímání zvukových podnětů, mezi něž patři i hlas, je nelineární. Při analýze hlasu je třeba tuto skutečnost brát v potaz a použít vhodnou metodu zaměřenou na nižší hladiny frekvencí. Metoda Mel-frekvenčních kepstrálních koeficientů pro kompenzaci nelinearity využívá tzv. banku filtrů. Počet filtrů obsažených v bance je úměrný vzorkovací frekvenci.

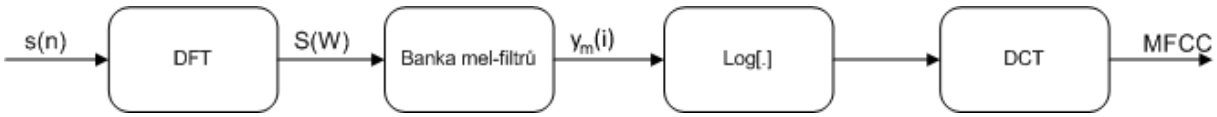
Tabulka 3.2: Počet pásem filtru

Vzorkovací frekvence [Hz]	*Počet pásem*
8 000	*15*
11 000	*17*
16 000	*20*
22 000	*22*
44 000	*27*

Jednotlivé filtry se překrývají v polovině svých pásem. Znamená to obdobně jako u segmentů při segmentaci, že každý následující filtr má počátek pásma v polovině pásma filtru předešlého. Šířka pásem jednotlivých filtrů je stejně velké pro nelineární melovskou stupnici a pro lineární frekvenční stupnici se šířka pásma filtrů zvětšuje pro vyšší frekvence. Vztah 3.13 definuje převodní vztah mezi frekvencí v Hertzech a v Melech.

$$f_{mel} = 2595 \log\left(1 + \frac{f}{100}\right) \tag{3.13}$$

Dle schématu na obrázku 3.11 je nejprve každý segment analyzovaného signálu převeden z časové oblasti do frekvenční oblasti aplikováním diskrétní Fourierovy transformace. Spektrum signálu je následně vynásobeno filtry. Dále je třeba všechny hodnoty v rozmezí intervalu oken filtrů sečíst a umocněním získat energii. Energie je následně zlogaritmována a prostřednictvím diskrétní kosinovy transformace získáme mel-frekvenční kepstrální koeficienty. [11]



Obrázek 3.11: Digram průběhu MFCC analýzy

3.3 Databáze hlasových vzorků

Pro vyhodnocení základního tónu řeči a srovnání kvality jednotlivých metod je použita dvojice databází hlasů. První balíček je berlínská databáze složená z více než 500 vybraných kvalitních nahrávek v německém jazyce se vzorkovací frekvencí 16 kHz. Jednotlivé záznamy jsou uloženy ve formátu WAV. Nahrávky jsou rozděleny systematicky podle čísla řečníka, čísla vyslovované věty a emoce, která byla v dané nahrávce simulována.

03a02Nc.wav

Obrázek 3.12: Význam označení nahrávky

Tabulka 3.3: Význam zkratek

Značka	*Označení emoce německy*	*Označení emoce česky*
A	*Angst*	*Strach*
E	*Ekel*	*Znechucení*
F	*Freude*	*Štěstí*
L	*Langeweile*	*Nuda*
N	*Neutral*	*Neutrální*
T	*Treuer*	*Smutek*
W	*Wut*	*Zlost*

Na obr. 3.12 je znázorněn význam formátování názvu nahrávek. První, zelené dvojčíslí označuje číslo mluvčího. Druhá, oranžová dvojice čísel nese označení mluvené věty a poslední modré písmeno značí emoční stav použitý v dané větě. V tabulce 3.3 jsou rozloženy všechna označení jednotlivých emocí v českém a německém jazyce spolu s přiřazeným.

Druhá databáze je složena ze záznamů zvuků delších větných celků, které byly následně rozstříhány na kratší zvukové úseky. Pořízené nahrávky jsou opět ve formátu WAV, přičemž jejich vzorkovací frekvence je nižší a to 8 kHz. Kvalita zvuku souborů je znatelně horší než u berlínské databáze.

4 Praktické vypracování

Dle zadání bakalářské práce je náplní praktické části vypracování algoritmu pro extrakci základní frekvence z hlasového vzorku. Proces extrakce je proveden dvěma metodami, a to prostřednictvím kepstrálních koeficientů a Mel-frekvenčních kepstrálních koeficientů. Po konzultaci s vedoucím bakalářské práce jsem se rozhodl praktické vypracování doplnit o třetí metodu, tedy extrakci pomocí Autokorelační funkce. Celý algoritmus je reprezentován, jako zdrojový kód pro matematické prostředí Matlab R2012a. V následujících kapitolách je funkce celého algoritmu vysvětlena pomocí diagramu. Všechny naměřené hodnoty jsou názorně zapsány v tabulkách a na závěr vyhodnocena jejich efektivnost.

4.1 Hlavní funkce fundamental_frequency

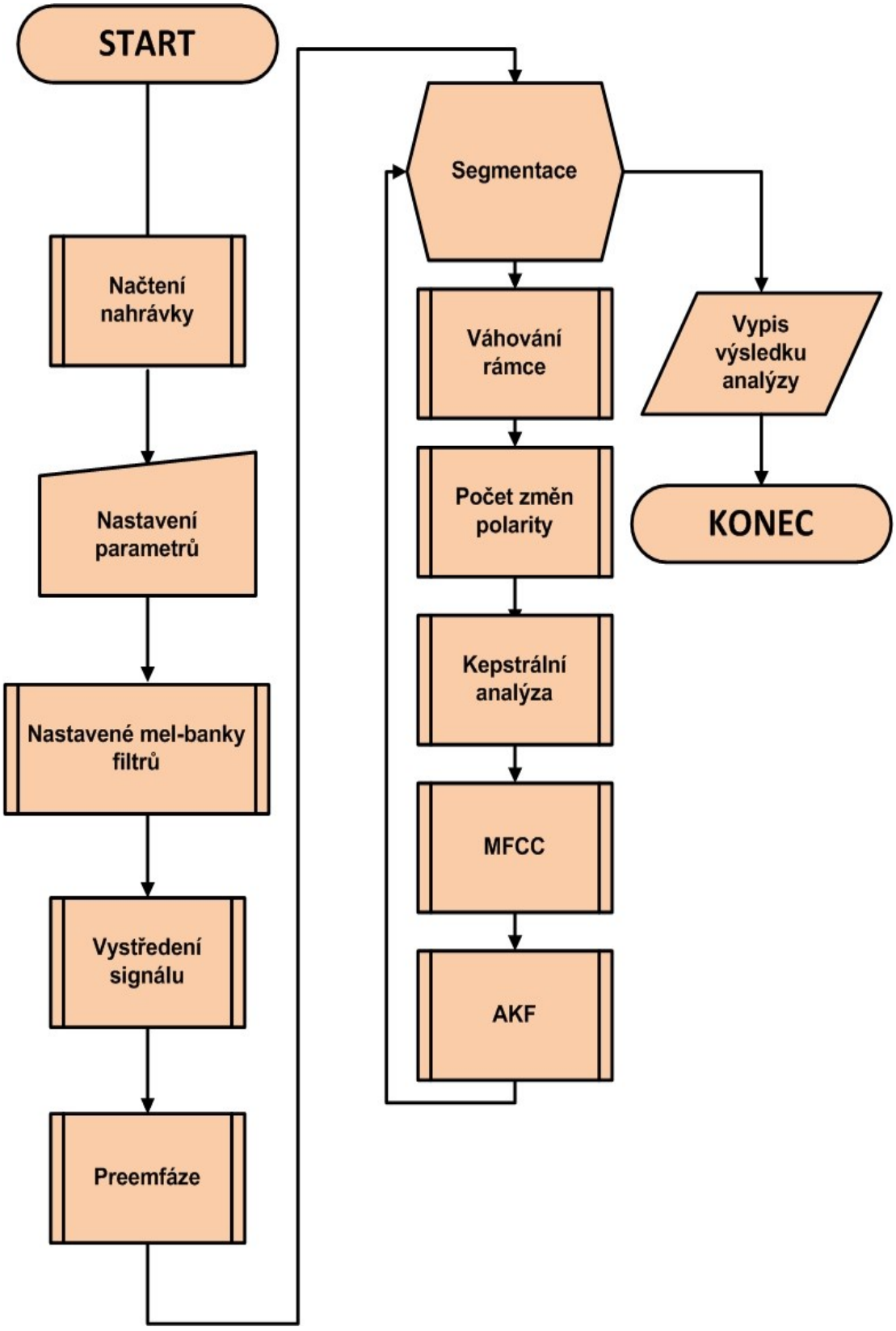
Tato funkce je hlavní funkcí skriptu. Nachází se v ní celý algoritmus analýzy řečového vzorku pro všechny tři metody. Pro použití této funkce v jiných skriptech není třeba na vstup přivádět žádná data. Funkce má svůj interní vstup pro data, tedy zvukovou nahrávku pro analýzu. Na výstupu algoritmu pak jsou tři proměnné obsahující každá jednu hodnotu základní frekvence pro právě jednu metodu. Celý algoritmus hlavní funkce je graficky znázorněn na obrázku 4.1 v podobě diagramu.

4.1.1 Nastavení základních parametrů

Na začátku kódu jsou stanoveny základní parametry (viz tabulka 4.1). Velikost rámce je pevně stanovena na 20 ms. Každý rámec musí začínat v polovině předcházejícího rámce (kapitola 3.1.4), proto je tedy překrytí rámce rovno 10 ms. Pro proces segmentace je nejpodstatnější zjistit počet rámců, aby mohl být signál přesně a pravidelně rozdělen do jednotlivých segmentů a proto je vhodnější převést proměnné **l_frame** a **o_frame** na proměnné **s_frame** a **o_frame_s**. Následně jsou jejich hodnoty dosazeny do vztahu 3.4 pro výpočet počtu rámců pro daný signál.

Tabulka 4.1: Základní parametry

	název proměnné
velikost rámce v sekundách	*l_frame*
překrytí rámce v sekundách	*o_frame*
počet vzorků v signálu	*l_signal*
počet vzorků v rámci	*s_frame*
počet vzorků překrytí	*o_frame_s*
začátek rámce	*frame_start*
konec rámce	*frame_end*
počet rámců	*noFrames*



Obrázek 4.1: Diagram funkce fundamental_frequency

4.1.2 Načtení zvukové nahrávky

Aby bylo vůbec z čeho extrahovat základní frekvenci, je třeba na začátku načíst vzorek hlasu pro testování.

```
[nazev,cesta] = uigetfile('*.wav','Vyberte audio nahrávku');
```

Pro navzorkování nahrávky je nejprve nezbytné zadat, kde se vůbec daný vzorek nachází a získat úplnou cestu k němu. Příkaz **uigetfile** otevře grafické rozhraní pomocí něhož snadno lze najít umístění vzorku. Na výstupu tohoto příkazu je uložen zvlášť název souboru, jako **nazev** a cesta k tomuto souboru, jako **cesta**. Příkaz také pro úplnost, za název souboru připojí příponu **.wav**. Je nutné podotknou, že zvuková nahrávka musí být uložena ve formátu s příponou **.wav** a musí být jednokanálová, tedy mono. Jinak by mohlo dojít k chybnému vyhodnocení celé analýzy nebo by vůbec nemusela proběhnout.

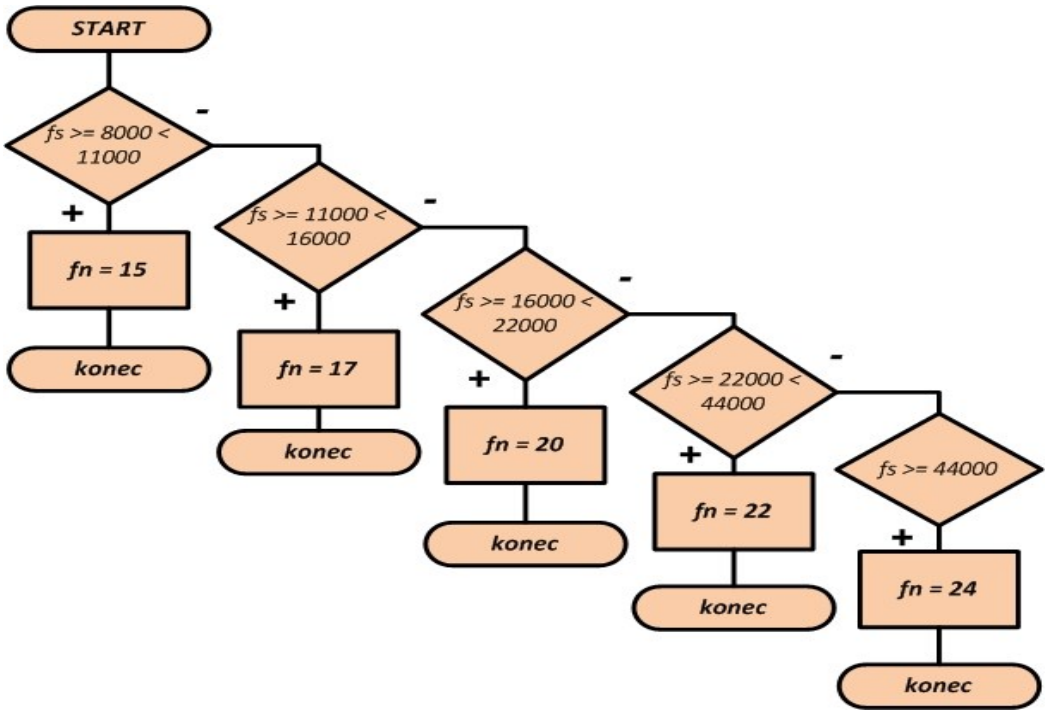
```
n = fullfile(cesta, nazev);
```

Následně je příkazem **fullfile** zkompletována cesta k souboru spolu s názvem nahrávky a na výstupu je celý řetězec znaků uložen, jako **n**.

```
[s fs] = wavread(n);
```

V dalším kroku funkce **wavread** načte nahrávku. Na výstupu jsou do proměnné **s** uloženy jednotlivé hodnoty vzorků seřazené podle času od začátku nahrávky až po její konec, jako vektor hodnot. Proměnná **fs** obsahuje hodnotu vzorkovací frekvence nahrávky.

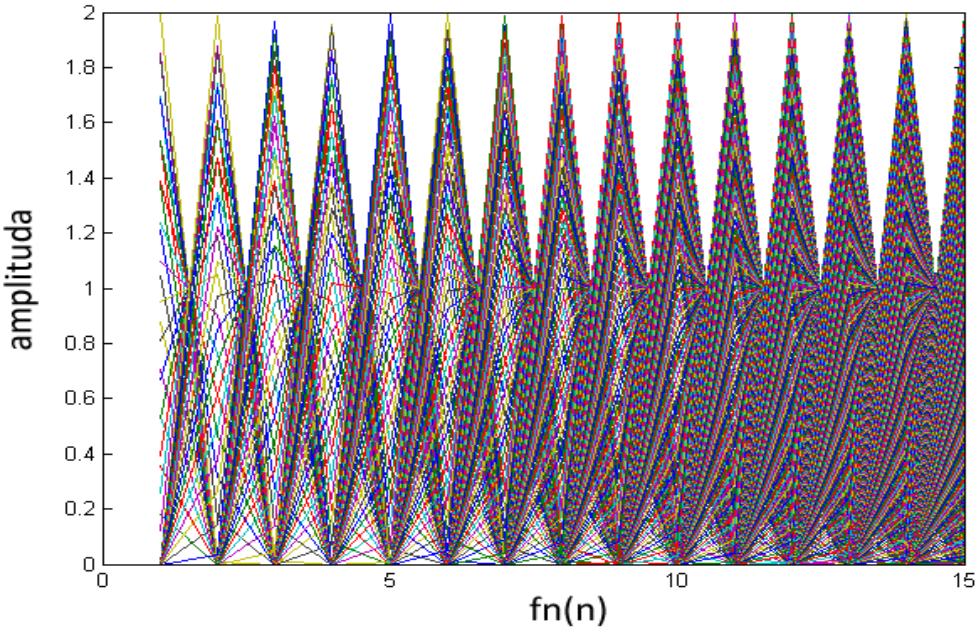
4.1.3 Nastavení banky melfiltrů



Obrázek 4.2: Digram volby počtu filtrů

Na obrázku 4.2 je znázorněn diagram algoritmu nastavení počtu filtrů v bance mel-filtrů. Na vstupu je porovnána vzorkovací frekvence nahrávky **fs**. Podle její velikosti je následně vyhodnocen počet filtrů dle tabulky 3.2.

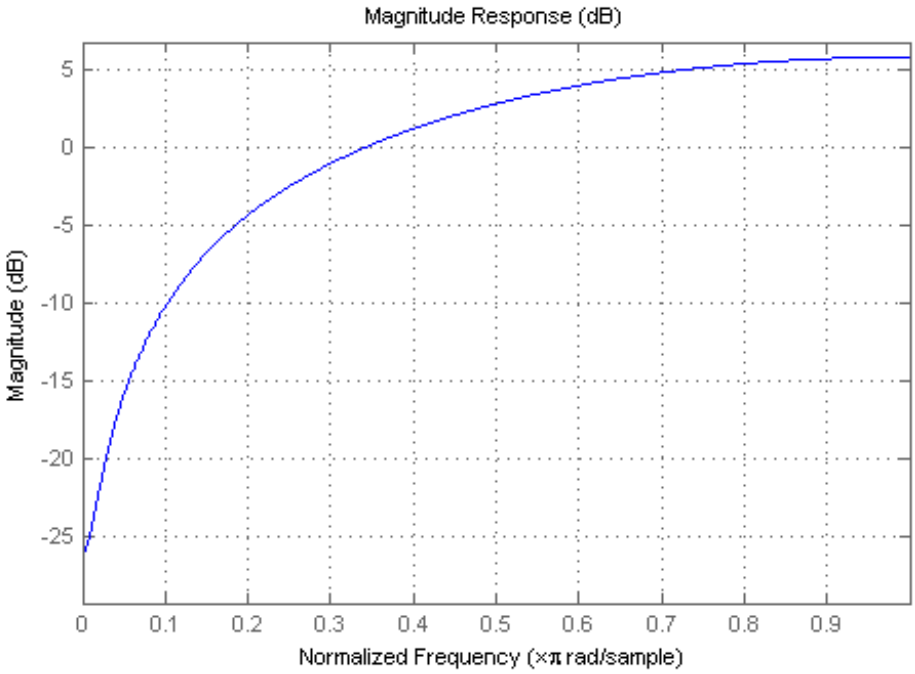
Na obrázku 4.3 je vykreslen průběh banky filtrů pro nahrávku se vzorkovací frekvencí o velikosti 8000 Hz. Banka tedy skýtá 15 filtrů.



Obrázek 4.3: filtry v bance Mel-filtrů

Algoritmus pro vytvoření banky mel-filtrů je volán, jako podfunkce ze souboru **melbankm.m**.

4.1.4 Preemfáze



Obrázek 4.4: Průběh premefázového filtru

Filtr, kterým je signál filtrován pro vyrovnání útlumu (viz 3.1.3) je vykreslen na obr. 4.4. Každý vzorek signálu po průchodu filtrem je uložen do proměnné **pre_s**.

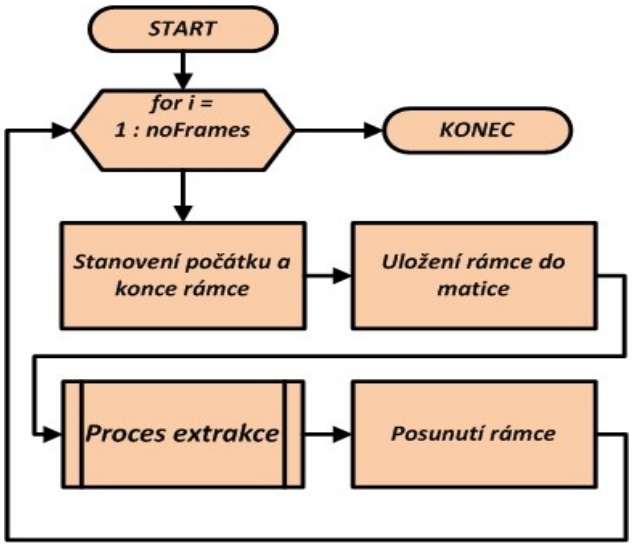
4.1.5 Cyklus segmentace

Struktura cyklu vychází z teorie v kap. 3.1.4. Počet průchodů cyklu je stanoven v rozmezí od prvního po poslední rámec. Přičemž počet rámců je stanoven na začátku funkce (viz kap. 4.1.1), stejně jako začátek a konec rámce.

```
frame = pre_s(frame_start:frame_end);

smatrix(:,i) = frame;
```

Z proměnné **pre_s** jsou vybrány vzorky po preemfázi z intervalu stanoveného pro aktuální rámec. Index prvního vzorku reprezentuje proměnná **frame_start** a posledního pak **frame_end**. Následně proběhne analýza (viz obrázek 4.5) a rámec je posunut přičtením hodnoty **o_frame_s**.



Obrázek 4.5: Digram procesu segmentace

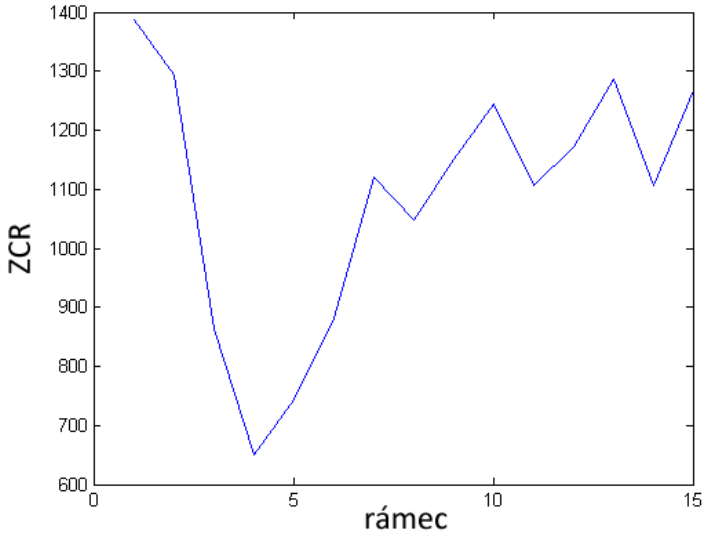
Jak již bylo řečeno v kapitole 4.1.2 pro více kanálové nahrávky, by nebylo možné provést analýzu vzorku. Algoritmus segmentace dokáže pracovat s hodnotami z jednosloupcové matice, kde jsou jednotlivé vzorky ukládány v čase sestupně. Pro dva a více kanálové zvukové záznamy je vytvářena matice více než jednosloupcová. Tento algoritmus tedy pracuje pouze s jedním sloupcem a zbylé hodnoty ostatních kanálů jsou vynechány. Analýza jednokanálových nahrávek je pro vyhodnocení jednotlivých metod postačující a proto by bylo zbytečné vytvářet složitější a robustnější algoritmus pro analýzu kvalitnějších nahrávek.

4.1.6 Váhování rámce a výpočet počtu změn polarity

Funkce pro váhování rámce Hammingovým oknem vychází z kapitoly 3.1.5 kdy je každý rámec vynásoben Hammingovým oknem pro zahlazení okrajů přechodů. Charakter okna **H** je stanovena funkcí **hamming** v Matlabu. Jako vstupní hodnotou udávající šířku okna je proměnná **s_frame**. Udává počet vzorků v rámci. Jiná šířka okna nepřipadá v úvahu. Jelikož musí být vždy násobeny matice o stejné velikosti. Okno musí být tedy stejně velké, jako sám rámec.

Výpočet energie (viz 3.2.1) a ZCR (viz 3.2.2) opět vychází z teorie. Hodnoty těchto dvou funkcí jsou podstatné, především při rozhodování, zdali je daný rámec znělí nebo neznělí. Jiný podstatnější význam nemají.

Obrázek 4.6 ukazuje počet přechodů přes nulovou osu v jednotlivých rámcích. S klesající hodnotou **ZCR** klesá i hodnota energie signálu a právě v této oblasti je základní frekvence nejmenší a lze předpokládat i přítomnost neznělého rámce.



Obrázek 4.6: Průběh ZCR

4.1.7 Kepstrální, MFCC a AKF analýza

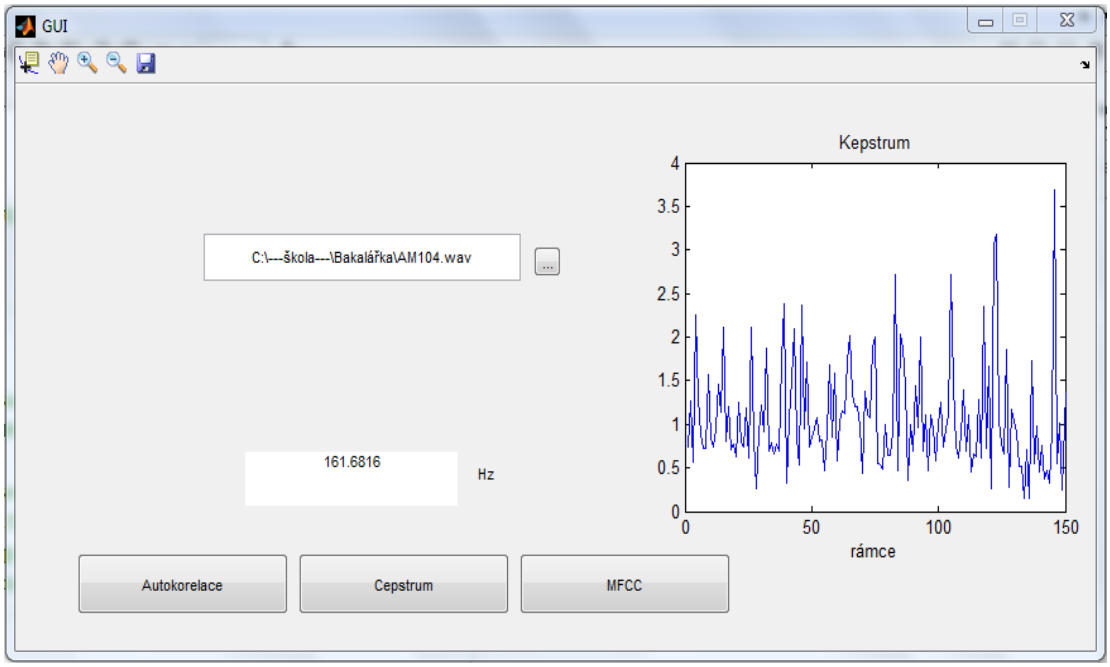
Na konci celého cyklu segmentace přechází rámec do oblasti vlastní analýzy. Na rámec po váhování je aplikována funkce **cceps**. Je to funkce Matlabu za níž se ukrývá celý proces kepstrální analýzy (viz 3.2.5). Na výstupu funkce jsou do matice **cepstrum** ukládány zprůměrované hodnoty kepstrálních koeficientů analyzovaného rámce. Pro získání základní frekvence celého signálu je třeba absolutní hodnoty matice sumarizovat. Což je provedeno na konci programu.

Implementace analýzy pomocí Mel-frekvenčních kepstrálních koeficientů je o něco složitější, jelikož pro ni neexistuje příslušná předpřipravená funkce. Jednotlivé kroky vychází z teorie kap. 3.2.6. Rámec je Fourierovou transformací převeden do frekvenční oblasti. Je zde určen počet transformačních koeficientů rovnající se počtu vzorků v rámci. Následně je vytvořena manka filtrů příslušném počtu filtrů (viz kapitola 4.1.3). Pro filtrování a následné vyhodnocení je brána v potaz jen první polovina výstupu transformace zbytek obsahuje nepodstatné informace pro nalezení $\mathbf{F_0}$. Po filtraci je rámec převeden do stupnice v Melech a je hledáno jeho maximum obsahující hodnotu $\mathbf{F_0}$. Výsledné hodnoty jsou opět vkládány do matice a sumarizovány. Tímto je extrahován základní tón metodou MFCC.

Extrakce prostřednictvím autokorelační funkce je podstatně jednodušší na výpočet než předešlé dvě metody. Na vstupu je matice segmentů vložena do pomocné proměnné **x** pro lepší práci s ní. Funkcí **xcorr** je provedena autokorelace rámce. Následně je hledána prahová úroveň rámce dle vztahu 3.8. Pokud je tato hodnota menší nebo rovna hodnotě **thr** je daný rámec neznělý. V tomto případě nastává jeden z případů, kdy je výsledná hodnota nepodstatná a je přepsána na nulu. Druhým případem je fakt, že je hodnota nekonečně velká. V tomto případě není možné dále určit $\mathbf{F_0}$. Proto je opět uložena nulová hodnota. Průměr získaných hodnot je pak roven $\mathbf{F_0}$.

4.2 Grafické rozhraní

Pro lepší přehlednost a usnadnění manipulace s jednotlivými metodami je vhodnější použít grafické rozhraní. Stejně jako celý algoritmus extrakce základního tónu hlasu, je i toto prostředí navrženo a vytvořeno v programu Matlab.



Obrázek 4.7: Grafické rozraní

Rozhraní je reprezentováno jednoduchým oknem. Horní lišta obsahuje nástroje totožné s nástroji pro práci s grafickým výstupem v rozhraní **figure** v Matalbu. První ikona zprava, disketa, slouží pro uložení grafického výstupu. Pak ikony s lupami slouží pro přibližování či oddalování průběhu v okně. Nástroj s ikonou ruky umožňuje průběh posouvat v libovolném směru podle obou os. Poslední ikona ukrývá pomůcku pomocí níž lze nechat si zobrazit příslušnou hodnotu v libovolném místě průběhu signálu. Vypíše hodnoty na x-ové i y-ové ose v určeném místě. Okno níže pod lištou slouží pro vložení cesty k vzorku pro analyzování. Pokud není přesná cesta známá, lze jej vyhledat pomocí tlačítka vedle něj. To otevře další grafické okno, pomocí něhož vzorek lze vyhledat a následně je uložena úplná cesta k nahrávce. Ve spodní části grafického rozhraní jsou patrná tři tlačítka, přičemž každé z nich nese název metody, která bude provedena po stisknutí. Následně se extrahovaná hodnota základního tónu vypíše v okně nad těmito tlačítky. Hodnota je uvedena v Hertzích. Výsledný průběh je vykreslen v okně na pravé straně rozhraní. Celkový vzhled grafického rozhraní je vyobrazen na obrázku 4.7. Celý zdrojový kód grafického rozhraní je spolu s návodem uložen na přiloženém CD/DVD, jako soubor graficke_rozhrani.fig. Pro jeho spuštění je potřeba program Matlab R2012b. Pokud nejde rozhraní spustit ze souboru graficke_rozhrani.fig, je třeba jej spustit pomocí souboru se zdrojovým kódem graficke_rozhrani.m.

5 Vyhodnocení jednotlivých metod

Poslední kapitola bakalářské práce je zaměřena na vyhodnocení výsledků získaných analýzou databáze zvukových nahrávek řeči. Pro analýzu je použita databáze (viz kapitola 3.3). Každý vzorek je analyzován zvlášť všemi třemi metodami a výsledné hodnoty v hertzích zapsány do tabulky. Berlínská databáze je složena z nahrávek hlasů 5 různých žen a 5 mužů.

5.1 Porovnání výsledků

V této kapitole jsou rozebrány extrahované základní frekvence pro jednotlivé emoční stavy nahrávek hlasu jednoho muže a jedné ženy. Následující zhodnocení jsou provedena porovnáním výsledných hodnot vůči emočním stavům jednotlivých nahrávek.

Tabulka 5.1: Muž - vztek

název	*CC*	*MFCC*	*AKF*
03a04Ad	*86,953*	*59,1315*	*60,3175*
03a05Aa	*159,8052*	*35,8916*	*0,81421*
03b02Aa	*127,7159*	*65,5271*	*60,4444*
03b10Ab	*163,3444*	*47,4203*	*34,5529*
průměr	*134,4546*	*51,99263*	*39,03225*

Tabulka 5.2: Žena - vztek

název	*CC*	*MFCC*	*AKF*
08a01Ab	*123,2323*	*57,5871*	*58,6986*
08a02Ab	*108,2326*	*94,106*	*62,8571*
08a02Ac	*77,152*	*9,5483*	*40*
08b01Aa	*135,4498*	*48,6109*	*31,4797*
08b09Ab	*150,359*	*55,7336*	*40,6407*
08b10Aa	*125,5814*	*7,5509*	*27,6923*
průměr	*120,0012*	*45,5228*	*43,5614*

Prvním emočním stavem je vztek. Z tabulek 5.1 a 5.2 je patrné, že v této části analýzy dosahuje nejlepších výsledků metoda kepstrálních koeficientů. Dle naměřené frekvence lze usoudit, že nahraný ženský hlas je v porovnání s mužským hlasem o něco hrubší.

Tabulka 5.3: Muž - štěstí

název	*CC*	*MFCC*	*AKF*
03a01Fa	*89,8543*	*51,9452*	*106,6667*
03a02Fc	*135,6082*	*78,2994*	*49,7364*
03a04Fd	*121,2239*	*103,7148*	*140*
03a05Fc	*183,501*	*33,6804*	*31,6619*
03a07Fa	*110,0771*	*30,3077*	*49,7011*
03b01Fa	*135,9138*	*40,4845*	*31,4604*
průměr	*129,3631*	*56,40533*	*68,20442*

Při extrakci základní frekvence z mužského hlasu (viz tab. 5.3) opět nejlepších hodnot dosahuje kepstrální analýza. Hodnoty ostatních dvou metod jsou pro většinu nahrávek příliš nízké, ale u třetího vzorku je reálná nejen hodnota pro CC, ale i pro autokorelační funkci.

Tabulka 5.4: Žena - štěstí

název	*CC*	*MFCC*	*AKF*
08a01Fd	*89,3834*		*43,8961*
08a02Fe	*76,8685*	*12,278*	*70,2632*
08a04Ff	*105,1234*	*79,3905*	*89,1393*
08a05Fe	*169,8978*	*115,8979*	*75,4674*
08a07Fd	*94,7573*	*122,439*	*37,0399*
08b01Fd	*94,8319*	*24,6578*	*40,4319*
08b01Fe	*127,6684*	*106,5082*	*68,3501*
08b02Ff	*160,9564*	*132,4844*	*70,0412*
08b03Fe	*225,8677*	*14,6175*	*98,2653*
08b09Fd	*158,2584*	*106,0201*	*36,9901*
08b10Fd	*127,4015*	*55,3276*	*113,4979*
průměr	*130,0922*	*76,9621*	*67,58022*

Podle vyhodnocených výsledků z tabulky 5.4 extrahovala autokorelační metoda velice zkreslené výsledky oproti kepstrální analýze, jejíž hodnoty lze považovat za nejkvalitnější. Ze zprůměrovaných hodnot z tabulek 5.1, 5.2, 5.3 a 5.4 je očividné, že emoce štěstí a vzteku dosahují obdobných hodnot $\mathbf{F_0}$. Základní frekvence obou hlasů je položena do stejné úrovně. Kvalitu kepstrální metody lze také vyvodit z podobnosti výsledků. Frekvence obou hlasů dosahují nejvyšších hodnot při vyslovování stejné věty (tj. č. 05). Není brána v potaz věta č. 03, jelikož pro mužský hlas nebyla analyzována a nebyla umístěna v databázi s mužským hlasem č. 03.

Tabulka 5.5: Muž - nuda

název	*CC*	*MFCC*	*AKF*
03a04Lc	*110,0232*	*95,4215*	*128*
03a07La	*106,4049*	*103,6859*	*109,3333*
03b01Lb	*90,0417*	*84,362*	*92,2873*
03b02La	*171,954*	*93,5744*	*113,0136*
03b09La	*176,7071*	*88,1144*	*120*
průměr	*131,0262*	*93,0316*	*112,5268*

Tabulka 5.6: Žena - nuda

název	*AKF*	*CC*	*MFCC*
08a01Lc	*80,7402*	*95,6811*	*17,7992*
08a02La	*36,3048*	*98,4249*	*53,1286*
08a04La	*81,2628*	*110,8293*	*183,9064*
08a05Lc	*26,154*	*179,315*	*38,0357*
08a07La	*78,8186*	*115,6546*	*46,8166*
08b01Lb	*33,8164*	*125,6952*	*32,0648*
08b02La	*105,3792*	*179,2943*	*51,3375*
08b03Lc	*81,6908*	*151,359*	
08b09Lc	*88,7668*	*150,0135*	*70,7592*
08b10La	*61,3993*	*131,5405*	*11,076*
průměr	*67,43329*	*133,7807*	*56,10267*

U této emoce jsou v tab. 5.5 hodnoty pro všechny nahrávky rovnoměrně rozložené. V průměru je nejvhodnější pro extrakci MFCC, jejíž klasifikace $\mathbf{F_0}$ je ve výsledku nejvěrohodnější vůči přisuzovanému emočnímu stavu. Naopak pro nahrávky ženského hlasu (viz tab. 5.6) je nejvhodnější AKF.

Tabulka 5.7: Muž neutrální

název	*CC*	*MFCC*	*AKF*
03a01Nc	*109,7281*	*106,3682*	*160*
03a02Nc	*43,5529*	*104,7256*	*91,4286*
03a04Nc	*100,7697*	*115,7994*	*120*
03a05Nd	*187,7461*	*100,1618*	*80*
03a07Nc	*92,0499*	*156,2182*	*89,4815*
03b01Nb	*154,1186*	*80,3956*	*96*
03b02Na	*166,8735*	*92,693*	*110,3111*
03b03Nb	*210,1854*	*71,6179*	*138,5185*
03b09Nc	*150,9625*	*73,0704*	*110,8887*
03b10Na	*107,7455*	*133,9387*	*87,5372*
03b10Nc	*150,1953*	*93,052*	*116,3636*
průměr	*133,9934*	*102,5492*	*109,139*

Tabulka 5.8: Žena - neutrální

název	*AKF*	*CC*	*MFCC*
08a01Na	*93,8005*	*81,0294*	
08a02Na	*71,3805*	*56,0765*	
08a04Nc	*96*	*97,5192*	*57,109*
08a05Nb	*63,7726*	*162,4038*	
08a07Na	*92,9524*	*78,8348*	
08b01Na	*75,9949*	*85,4854*	
08b02Nb	*92,707*	*158,8751*	
08b03Nb	*94,0942*	*198,9972*	
08b09Nb	*91,8974*	*134,2107*	*7,3441*
08b10Nc	*80,1826*	*123,1676*	*24,5343*
průměr	*85,27821*	*117,66*	*29,66247*

Neutrální emoce reprezentuje klasicky mluvené slovo, kdy není kladen žádný temperament. S ohledem na toto hledisko by se měla $\mathbf{F_0}$ pohybovat pod hranicí 100 Hz. Nejlepších hodnot dosahuje metoda MFCC pro mužský hlas (viz tabulka 5.7) a metoda AKF pro ženský hlas (viz tabulka 5.8), a to za předpokladu, že ženský hlas je posazený frekvenčně níže než analyzovaný mužský hlas.

Poslední testovanou sadou vzorků jsou nahrávky s emocí zlosti. Tyto záznamy mají dispozici pro nejvyšší základní frekvenci. Vzorky mužského hlasu (viz tabulka 5.9) jsou nejlépe vyhodnoceny metodou kepstrálních koeficientů, stejně je tomu tak i pro banku ženských vzorků (viz tabulka 5.10). V některých případech je lepší pro další potřebu zpracování vyhodnotit základní frekvenci jako průměrnou hodnotu stanovenou všemi třemi metodami nebo vybrat nejlepší vhodnou hodnotu $\mathbf{F_0}$. Jak je vidět např. v tabulkách 5.6 a 5.8, někdy dochází k absolutně špatnému vyhodnocení základní frekvence, a proto není tato hodnota uvedena v tabulce.

Tabulka 5.9: Muž - zlost

název	*CC*	*MFCC*	*AKF*
03a01Wa	*107,1724*	*103,5186*	*125,9259*
03a02Wb	*132,4742*	*96,8594*	*34,1818*
03a02Wc	*88,7406*	*160,9541*	*48,3883*
03a04Wc	*135,091*	*141,468*	*96,5818*
03a05Wa	*187,7982*	*25,9737*	*75,2941*
03a05Wb	*208,1014*	*135,2751*	*75,2653*
03a07Wc	*150,122*	*68,3037*	*55,2*
03b01Wa	*124,2914*	*46,2822*	*58,3784*
03b01Wc	*148,5168*	*23,7635*	*53,6416*
03b02Wb	*191,5641*	*90,7531*	*100*
03b03Wc	*251,5508*	*34,8408*	*67,5042*
03b09Wa	*160,7625*	*39,8606*	*72,8835*
03b10Wb	*143,633*	*74,3888*	*76,1133*
03b10Wc	*160,1318*	*152,3266*	*94,3218*
průměr	*156,425*	*85,3263*	*73,83429*

Tabulka 5.10: Žena - zlost

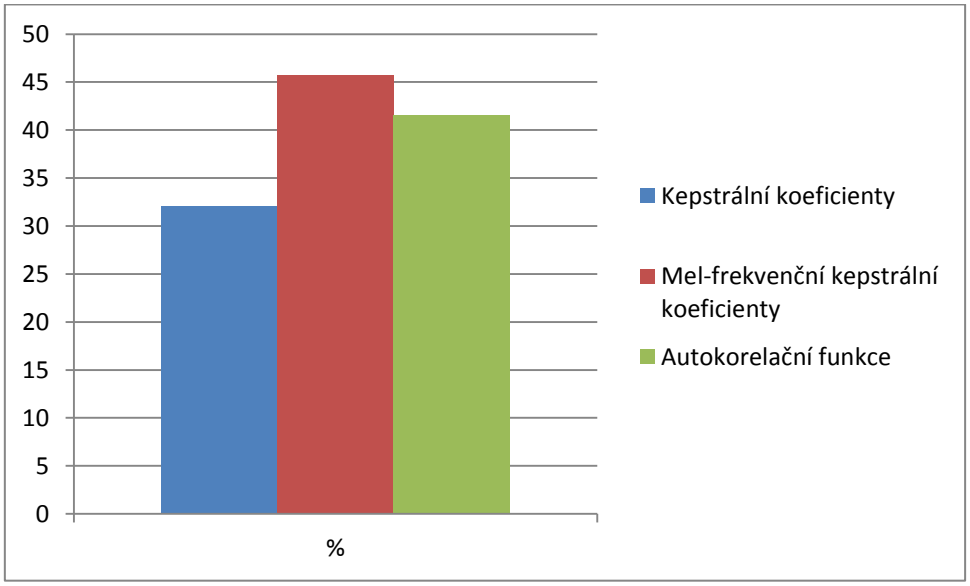
název	*AKF*	*CC*	*MFCC*
08a01Wa	*3,6043*	*104,5182*	*117,729*
08a01Wc	*1,0283*	*86,3846*	*194,4342*
08a02Wc	*88,8889*	*104,7022*	*147,5321*
08a04Wc	*88,8889*	*104,7022*	*147,5321*
08a05Wa	*46,8113*	*159,9528*	*131,8987*
08a07Wc	*37,4667*	*145,9361*	*23,215*
08b01Wa	*56,6627*	*202,1057*	*56,491*
08b02Wd	*57,7778*	*159,2075*	*51,8475*
08b03Wd	*101,4286*	*269,331*	*8,0142*
08b09Wa	*41,801*	*187,0152*	*76,2044*
08b09Wc	*78,6704*	*208,6609*	*7,1412*
08b10Wa	*32,163*	*191,8789*	*54,2109*
průměr	*52,93266*	*160,3663*	*84,68753*

5.2 Statistické zhodnocení

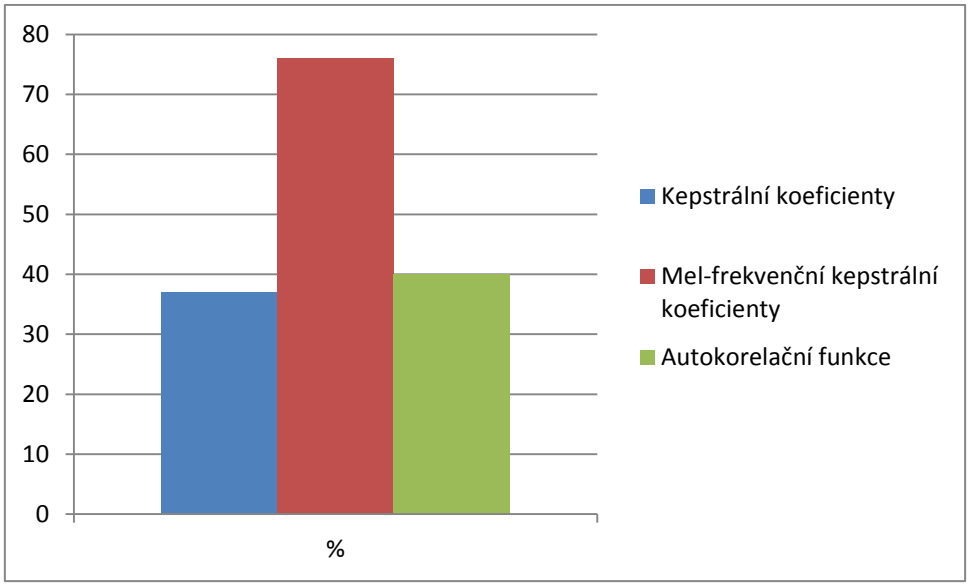
Z naměřených hodnot v tabulkách 5.1 až 5.10 je třeba jednoznačně určit metodu, jejíž výsledky jsou nejkvalitnější. Pro tento odhad je použito statistické vyhodnocení pomocí míry variability.

$$V_x = \frac{s}{\bar{x}} \cdot 100 \tag{5.1}$$

Dle vztahu 5.1 lze vypočítat míru variability pro jednotlivé metody analýzy. Na výstupu je získána hodnota homogenity extrahovaných hodnot v procentech. Variační koeficient vyšší než 50% je známkou vysoké nehomogenity dané metody. Jednotlivé výsledky pocházejí ze širokého intervalu hodnot, což značí nepřesné vyhodnocování.



Obrázek 5.1: Statistické vyhodnocení metod pro analýzu mužského hlasu



Obrázek 5.2: Statistické vyhodnocení metod pro analýzu ženského hlasu

Grafy na obrázcích 5.1 a 5.2 zobrazují míru variability a její variační koeficient jednotlivých metod. Z analyzovaných hodnot mužského i ženského hlasu vyplývá, že metoda kepstrálních koeficientů poskytuje nejucelenější výsledky. Naopak metody Mel-frekvenční kepstrálních koeficientů a autokorelační funkce poskytují více nehomogenní informace. Při analýze ženského hlasu přesáhl variační koeficient metody MFCC hranici 76%. Výsledky ze statistického zhodnocení jednotlivých metod extrakce se shodují s odhadem výsledků z kapitoly 5.1. Jednotlivé výpočty míry variability a variačního koeficientu spolu s naměřenými výsledky jsou uloženy jako příloha na CD/DVD v souboru statistika.xls.

6 Závěr

Cílem této bakalářské práce byla analýza kepstra řečových vzorků. Pomocí algoritmu realizovaného v prostředí Matlab R2012a je extrahována základní frekvence hlasu z jednotlivých nahrávek pomocí vybraných metod. Použita je metoda analýzy kepstrálních koeficientů, Mel-frekvenčních kepstrálních koeficientů a autokorelační funkce.

Pro jednodušší práci s algoritmem a následné získávání hodnot základního tónu ze zvukových souborů bylo použito vytvořené grafické rozhraní. Tato aplikace je uložena na přiloženém CD/DVD spolu s kompletní berlínskou databází a návodem pro použití programu. Je nutno podotknou, že pro spuštění aplikace Rozhraní je třeba program Matlab R2012a.

Jednotlivé kroky programu jsou praktickou reprezentací matematických vztahů, které jsou teoreticky vysvětleny v kapitolách 3.1 a 3.2. Přičemž celý zdrojový kód algoritmu funkce fundamental_frequency (viz kap. 4.1) je okomentovaný a uložen na přiloženém CD/DVD.

V kapitolách 5.1 a 5.2 se nachází vyhodnocení použitých technik. Odhad v kapitole 5.1 je potvrzen statistickým zhodnocením v kapitole 5.2. Dle výpočtu variačních koeficientů je pro extrakci základního tónu lidského hlasu ze zvukových WAV souborů nejvhodnější metoda kepstrálních koeficientů, kde variační koeficient u analýzy mužského i ženského hlasu nepřesáhl 37%. Oproti tomu hodnoty získané metodami AKF a MFCC při analýze mužského hlasu jsou o něco více zkreslené. Pro AKF je to variační koeficient 41,5% a pro MFCC je roven 45,7%. Autokorelační funkce při analýze ženského hlasu dosahuje téměř stejné homogenity jako CC, přičemž MFCC přesahuje 76%. Nejvhodnější a nejpřesnější metodou je tedy analýza pomocí kepstrálních koeficientů. Tyto rozdíly mezi jednotlivými metoda jsou způsobeny koncepcí jednotlivých nahrávek v návaznosti na jejich kvalitu. Pokud by byl implementován podrobnější a robustnější algoritmus bylo by možné dosáhnout vyšší přesnosti a ucelenosti výsledků jednotlivých metod. Z vyhodnocení (viz kap. 5.2) lze poukázat na fakt, že metoda AKF je o mnoho jednodušší než CC, ale i přesto se jí dokáže vyrovnat. Oproti tomu koncepčně složitější a propracovanější metoda MFCC nedosáhla očekávaných kvalitních výsledků.

Funkci fundamental_frequency lze jednoduše implementovat do libovolných větších programů. Jelikož na vstup této funkce není třeba nic přímo vkládat je zcela nezávislá. Na výstupu poskytuje přepokládané hodnoty základní frekvence s kterými je možné dále pracovat bez žádné další nutnosti jejich úpravy, jelikož jsou již připraveny a převedeny na hertze.

Pokud by byly do algoritmu implementovány další metody pro extrakci F_0 bylo by možné provést širší analýzu jednotlivých metod a nalézt tak vhodnější způsoby vyhodnocování. Popřípadě by bylo možné nalézt vhodné kombinace těchto metod pro objektivnější získávání výsledků a vytvořit tak nové metody. Tyto metody by nebyly ovlivněné např. šumem v nahrávkách, nebo pokud by došlo k chybnému vyhodnocení základní frekvence, byly by schopny tuto chybu odhalit a kompenzovat.

Použitá literatura

[1] PSUTKA, Josef, Luděk MÜLLER, Jindřich MATOUŠEK a Vlasta RADOVÁ. Mluvíme s počítačem česky. Vyd. 1. Praha: Academia, 2006, 746 s. ISBN 80-200-1309-1.

[2] Suprasegmentálne javy- TEMPO A RYTMUS REČI, PRESTÁVKA, melodia. [online]. [cit. 2013-05-06]. Dostupné z: http://referaty.hladas.sk/referat.php/suprasegmentalne-javy--tempo-a-rytmus-reci--prestavka--melodia/23/15760

[3] WELLS, Christopher J. Technology UK: Telecommunications principles - Pulse code modulation (PCM). [online]. [cit. 2013-05-06]. Dostupné z: http://www.technologyuk.net/telecommunications/telecom_principles/pulse_code_modulation.shtml

[4] *Introduction To Speech Processing*. Oregon, 2010. Dostupné z: http://seminar.csee.ogi.edu/reu/speechsignalprocessing2.pdf

[5] *Introduction To Speech Processing*. Aalborg universitet. Dostupné z: http://kom.aau.dk/group/04gr742/pdf/framing_worksheet.pdf

[6] MARTIN, Philippe. *Intonation: Intonation's many functions*. University of Toronto. Dostupné z: http://www.semioticon.com/virtuals/talks/martin.htm

[7] GUTIERREZ-OSUNA, Ricardo. A&M UNIVERSITY. *Introduction to Speech Processing: Cepstral analysis*. Texas. Dostupné z: http://www.semioticon.com/virtuals/talks/martin.htm

[8] ANGUERA, Xavier. *Cepstral analysis*. Dostupné z: http://www.xavieranguera.com/tdp_2011/8-Cepstral-Analysis.pdf

[9] ATASSI, Hicham. Metody detekce základního tónu řeči. Brno, 2008.

[10] UHLÍŘ, Jan. *Technologie hlasových komunikací*. Vyd. 1. Praha: Nakladatelství ČVUT, 2007, vii, 276 s. ISBN 978-80-01-03888-8.

[11] Zpracování číslicových signálů: vzorkování, rekonstrukce, okénkové funkce, filtrace, IIR, FIR filtry a jejich návrh. A/D a D/A převodníky, jejich principy a parametry [online]. [cit. 2013-03-17]. Dostupné z: http://statnice.obrys.cz/index.php?title=Zpracov%C3%A1n%C3%AD_%C4%8D%C3%ADslicov%C3%BDch_sign%C3%A1l%C5%AF_-_vzorkov%C3%A1n%C3%AD,_rekonstrukce,_ok%C3%A9nkov%C3%A9_funkce,_filtrace,_IIR,_FIR_filtry_a_jejich_n%C3%A1vrh._A/D_a_D/A_p%C5%99evodn%C3%ADky,_jejich_principy_a_parametry.&redirect=no

[12] MACQUARIE UNIVERSITY. *Speech recognition* [online]. Sydney: Department of Computing, 2002 [cit. 2013-03-17]. Dostupné z: http://web.science.mq.edu.au/~cassidy/comp449/html/ch05.html

Seznam příloh

Součástí BP je CD/DVD:

- Databáze hlasů
- Fundamental_frequency.m
- Graficke_rozhrani.fig
- Graficke_rozhrani.m
- Melbankm.m
- Pitchcor.m
- Statistika.xls
- BP_VYC0015.pdf
- Navod_k_rozhrani.pdf

Příloha.A: *Zdrojový kód funkce fundamental_frequency*

```
function [f0_by_mfcc,f0_by_autocorelation,f0_by_cepstrum] =
fundamental_frequency()


clc; clear all;close; % vyčištění obrazovky příkazů a paměti
proměnných


%% načtení zvukové nahrávky

[nazev,cesta] = uigetfile('*.wav','Vyberte audio nahrávku');%
otevření grafického rozhraní a načtení cesty k nahrávce

g = fullfile(cesta, nazev);% zkompletování cesty do jednoho řetězce

[s fs] = wavread(g);% načtení a vzorkování zvukové nahrávky


%% nastavení parametrů


l_frame = 0.2; % délka rámce v sekundách

o_frame = l_frame/2; % velikost překrytí v sekundách

l_signal = length(s); % počet vzorků v signálu

s_frame = l_frame * fs; % počet vzorků v rámci

frame_start = 1; % začátek rámce

frame_end = l_frame * fs; %konec rámce

o_frame_s = o_frame*fs; % velikost překrytí ve vzorcích

noFrames = 1+floor(((l_signal-o_frame_s) / fs)/l_frame);

% počet rámců

smatrix = zeros(s_frame,noFrames);

% vytvoření marice pro uložení jednotlivých rámců

H = hamming(s_frame); % vytvoření Hammingova okana o velikosti rámce

E = 0; % alokování proměnné E - Energi

zcr = zeros(noFrames,1); % alokování matice pro ZCR

frame2 = zeros(s_frame,1); % alokování matice pro rámec(n-1)

n = s_frame/2;              % nastavení počtu prvků v FFT

LMIN = 1; LMAX = 159;

% hranice frekvencí, kdy je předpokládáno, že f0 rámce nebude větší
než 159 Hz
```

```
lags = zeros(1,noFrames);  % vytvoření matice lags a naplnění nulami
thr = 0.5;                 % nastavení hodnoty treshold
F0 = 0;                    % nastavení počáteční hodnoty pro F0

%% nastavení počtu filtrů v mel-bance filtrů
if fs == 8000 %pokud je vzorkovací frekvence v rozmezí od 8kHz do
    fn = 15;          % 11kHz, bude počet filtrů nastaven na 15

else if fs == 11000 % pokud je vzorkovací frekvence v rozmezí od
        fn = 17;         % 11kHz do 16kHz, počet filtrů je roven 17

    else if fs == 16000 % pokud je vzorkovací frekvence v rozmezí
            fn = 20;            % od 16kHz do 22kHz, počet filtrů je
                                % roven 20

        else if fs == 22000  % pokud je vzorkovací frekvence v
                fn = 22;    %rozmezí od 22kHz do 44kHz, počet filtrů
                                %počet filtrů je roven 22

     else if fs >= 44000 % pokud je vzorkovací frekvence větší než
                    fn = 24;    % 44kHz, počet filtrů je roven 24
                end
            end
        end
    end
end


%% vystředění signálu - DC-Offset
mi = 1/length(s)*sum(s);
% průměrování signálu (vypočítání Stejn. složky)
for i = 1:l_signal
    s(i) = s(i) - mi; % odečtení složky od signálu
```

```
end
 %% Preemfáze
a = 1;          % nastavení parametrů filtru
b = [1 -0.95];  % nastavení parametrů filtru
%fvtool(b,a);   % zobrazení průběhu filtru
pre_s = filter(b,a,s); %filtorvání signálu

for i =1:noFrames
   %% segmentace
   frame = pre_s(frame_start:frame_end); % vložení vzorků do rámce

   %% váhování rámce
   frame = frame.*H; %vynásobení rámce Hammingovýmm oknem
   smatrix(:,i) = frame; %uložení rámce do matice
   %% výpočet energie
   E_frame = sum(frame.^2); % výpočet Energie rámce
   E = E + E_frame;% výpočet celkové energie nahrávky

   %% počet změn polarity
   frame2(2:end) = frame (1:end-1);% stanovení rámce předcházejícího
                                   % rámce
   zcr(i) = 1/2 * sum(abs(sign(frame)- ... % výpočet změn polarity
   sign(frame2)))/(s_frame*0.5)*s_frame;
   %% Cepstrum
   cepstrum(:,i) = mean(cceps(frame))*100;
   % výpočet kepstrální funkce
   %% MFCC
   fft_frame = fft(frame,s_frame); % Fourierova transformace rámce
   melfiltr = melbankm (fn,n,s_frame); % stanovení mel-filtru
   halfn=1+floor(n/2);    % stanovení poloviny transformace
   spectr1=2595*log(1+(melfiltr*... %převedení na mely
       abs(fft_frame(1:halfn)).^2)/1000);
   spectr=max(spectr1(:),1e-22); %hledání maxima proměnné
```

```
    c=dct(spectr); % diskrétní kosinova transformace
    mfcc(:,i) =sum(c); % sumarizace koeficientů
    %% Autokorelace
     x = smatrix(:,i);   % uložení matice
     N=length(x);        % zjištění velikosti matice
    [L,R]=pitchcor(x,LMIN,LMAX,thr); % výpočet autokorelační funkce
     lags (i) = L;
     f0_pitch(i)=fs/L;  % výpočet vzdálenosti peaku s f0 od maxima
     if f0_pitch(i)==Inf     % pokud je f0 nekonečná
       f0_pitch(i)=0;        % pak je nastavena hodnota 0
     end
         F0(i)=f0_pitch(i)/100; % uložení F0 daného rámce


    %% posunutí rámce
    frame_start = frame_start + o_frame_s; % posunutí začátku rámce
    frame_end = frame_end + o_frame_s;     % posunutí konce rámce
end
     %% výsledky
     f0_by_mfcc = mean(mfcc)/10; % výpočet f0 z MFCC
     f0_by_autocorelation =mean(F0);  % výpočet f0 z AKF
     f0_by_cepstrum=sum(abs(cepstrum))*100; % výpočet f0
    display('Základní frekvence extrahovaná pomocí MFCC')
     f0_by_mfcc
     display('Základní frekvence extrahovaná pomocí AKF')
     f0_by_autocorelation
     display('Základní frekvence extrahovaná pomocí Kepstra')
     f0_by_cepstrum
```